Pattern Learning in Infants and Neural Networks

Michael Gasser (gasser@indiana.edu)
Eliana Colunga (ecolunga@indiana.edu)
Cognitive Science Program, Indiana University
Bloomington, Indiana 47405 USA

(To appear in P. Quinlan (Ed.) (2002). *Connectionist Models of Development*. Brighton: Psychology Press.)

Pattern Learning in Infants

Two recent sets of experiments have greatly expanded our understanding of the capacity of infants to detect regularities in patterns.

Saffran, Aslin, and Newport (1996) conducted a set of experiments to investigate the sensitivity of infants to the statistical properties of patterns. Eight-month-olds heard strings of syllables consisting of randomly concatenated three-syllable "words", sequences that never varied internally. Thus the transition probabilities within words were higher than between words. There were four different words consisting of twelve different syllables, and the subjects heard 180 words in all. Later the infants indicated, through differences in looking times, that they differentiated between these words and non-word three-syllable sequences which they had either heard with less frequency than the words (because they consisted of sequences of syllables which crossed word boundaries) or not heard at all. This is taken as evidence that they had picked up the statistics in the training set. More recent experiments (Saffran et al., 1999) have achieved similar results with patterns consisting of tones of different pitches.

Marcus, Vijayan, Bandi Rao, and Vishton (1999) conducted a set of experiments to determine whether infants could extract rules from a pattern. Seven-month-olds were presented with series of three-syllable sequences separated by gaps. Each sequence consisted of two different syllables arranged in a fixed pattern, (*x x y*), (*x y y*), or (*x y x*). For example, in the (*x y y*) condition, the presented patterns included sequences such as *le di di* and *ji je je*. During the familiarization phase, each of 16 sequences was presented three times. Later the infants were tested on sequences consisting of novel syllables, in one experiment, syllables designed so as not to overlap with the training syllables. The test sequences either obeyed the training rule, or they obeyed one of the other two rules. The infants indicated, through differences in looking times, that they differentiated the patterns that followed the training rule from those that did not. This is taken as evidence that they had in some sense picked up the rule implicit in the training patterns.The experiments of Marcus et al. have attracted a good deal of attention because the authors, as well as Pinker (1999), have seen them as a challenge to connectionist models. Because connectionist models can only respond on the basis of similarity to training items, they apparently could not generalize to test sequences consisting of novel syllables, as the infants did. We believe that the line is not a simple one separating symbolic and connectionist models. The question is: what minimal set of mechanisms is required to achieve a behaviour, and what sorts of predictions a model embodying these mechanisms makes. In this paper we present an account of patterns and pattern learning that brings together the two sorts of experiments in a single framework, and we propose a minimal set of mechanisms based on this account that could achieve the sort of pattern learning we see in the experiments. We also show how a connectionist implementation of these mechanisms models the results of the experiments and makes novel predictions. This is not the first connectionist model of the Marcus et al. results (see, for example, Christiansen, Conway, & Curtin, 2000; Seidenberg et al., 1999; Shultz & Bale, 2000); however, we believe that the model we propose has the advantage of simplicity and perspicuity.[1]

Patterns

The world out there is full of regularities of one sort or another, and it is to our advantage to be able to find them. Much of the regularity takes the form of what we will call **patterns**, in which

[1] It is in fact considerably simpler than the model we previously proposed (Gasser & Colunga, 2000).

the regularity consists of recurring relations between the elements of a sequence, presented in time or in space.

For example, say you are presented with a sequence of syllables from a language unknown to you. If you listen to such a sequence long enough, you may be able to detect certain regularities: certain syllables tend to be followed by certain other syllables or tend not to be followed by certain other syllables; certain syllables tend to be repeated; the syllables tend to group together in relatively predictable sub-sequences.

Now consider a static, visual example. Each morning you are served breakfast. There are always two or three food items arranged in a row in front of you. Over time you notice that the juice is always to the left of the cereal and that if there is toast, it is always on the right.

There is no question that people (and other animals) have the ability to find regularity in patterns. But there are still questions about what they can and cannot discover, as well as about what regularities are out there to be discovered. And there are even more fundamental questions about what sort of mechanism is behind the learning and processing of patterns. Figuring all this out is of great importance because of the role patterns play in our lives. Language in particular is at least to a large extent a matter of patterns. In this paper, we consider what is involved in pattern learning, discuss two experiments with infants demonstrating pattern learning, and propose a connectionist model which accounts for the infant data.

Elements and Dimensions

The processing of patterns assumes a pre-processing stage in which the input stream has already been segmented into a sequence of **elements**, objects all belonging to a particular general category such as syllable, musical tone, or polygon.

Before they can discover the regularities in a given pattern, we also assume that people represent the elements in terms of values on one or more dimensions such as pitch or color. In addition, they may also be able to place the elements in one or another of a set of disjoint categories.

Between-Element Relations

The regularity within a pattern is defined in terms of relations between pattern elements. It is important to note that there may be more than one way to characterize the regularity in a given set of patterns. For human subjects, we of course do not have direct access to the manner in which they are representing the regularity, and there will normally be more than one underlying representation consistent with their performance on test patterns.

Because in the general case the relations between elements emerge out of tendencies present in multiple examples, we will speak of the relations as **correlations**.

There are two sorts of between-element correlations that may characterize patterns. In the simpler case, the relation is between specific element contents. When there are element categories, the correlation may between specific types. For example, the syllable *ba* may tend to be followed by the syllable *gu*; the cereal may tend to be on the right of the juice. Alternatively the correlation may be between specific values or ranges of values along element dimensions. For example, a syllable with the vowel *e* may tend to be followed by a syllable with the vowel *u*; a breakfast food item in a plastic bowl may tend to appear to the right of a hot drink. We will refer to correlations that make reference to the specific element categories or dimension values as **content-specific correlations**.

A second kind of correlation is more abstract; it concerns whether two elements *match* with respect to their category or a particular dimension. For example, in a sequence of syllables there may be a tendency for syllables to appear in same-category pairs: *bu bu go go ta ta…* Or adjacent syllables may tend to match on a given consonant or vowel feature such as rounding, a process

known as "harmony" in phonology. At the breakfast table, the three items may tend to appear in containers of the same color. We will refer to correlations of this type as **relational correlations** because their definition makes reference to the primitive relations SAMENESS and DIFFERENCE.

So far the correlations we have discussed relate pairs of elements. **Higher-order correlations**, correlations between pairwise co-occurrences, are also possible. Examples of content-specific higher-order correlations are the following:

- 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*.

Examples of relational higher-order correlations are the following:

- a pair of syllables of one type tends to be followed by a pair of another type
- when a syllable is preceded by another syllable beginning with the same consonant, it tends to be followed by a syllable ending in the same vowel
- when the consonants of two adjacent syllables match in one feature, they and the vowel in the syllables tend to match on all features.

Note that higher-order correlations do not presuppose that the co-occurrences they relate are actually correlations. In the first example of a content-specific correlation above, *gu* need not tend to be preceded by *ba* or tend to be followed by *li* for the correlation between these two co-occurrences to occur.

Groups and Segmentation

The regularities within patterns also break the sequences of elements into subsequences which we will call **groups**.

Groups may be distinguishable in four different ways (or some combination of these), all of which may contribute to segmentation of sequences of elements into groups.

1. The intervals separating groups may be distinguishable from the intervals separating elements within groups. In the breakfast example, the distinction is obvious. The elements within a group are separated from one another spatially, whereas the groups are separated from one another temporally. In other cases, the gaps between groups may simply be longer than the gaps between elements.
2. There may be similarity between elements in corresponding positions within groups. For example, in a sequence of tones, each group might begin on a relatively high pitch.
3. There may be stronger correlations within groups than between groups. This appears to be a fundamental property of natural language, for example. As more and more segments of a word become available, the remainder of the word becomes more and more predictable, but predictability across word boundaries, except in the case of idiomatic constructions, tends to be relatively low.
4. The within-group relations between corresponding elements may be similar. For example, the first two elements in each group might be tokens of the same type.

Pattern Behavior

What does it mean to know a pattern? There are two behaviours that can act as measure of pattern knowledge. One is the ability to distinguish acceptable sequences of elements, those that agree with the regularities in the pattern, from unacceptable sequences. Of course acceptability may be a matter of degree. Another is the ability to complete patterns, to fill in one part of a sequence given another. For a temporal pattern learning task, pattern completion means prediction: given one or more of the elements in a group, one or more subsequent elements are predicted.

Pattern Learning

We assume that patterns are learned through the presentation of examples to the learner. Thus pattern learning is an **unsupervised** task; a learning trial consists simply of a pattern subsequence, not of an input together with a target output. At any point during learning, the learner may be tested by being presented with a test sequence and expected to judge its acceptability or with a partial sequence and expected to complete it.

Many (but not all) pattern learning tasks are induction tasks; that is, they involve generalization from the set of training groups of elements to a larger, "target" set of acceptable groups. The traditional characterization of induction assumes that the learner maintains a "hypothesis" about what constitutes the target set, updating the hypothesis when it fails to accommodate particular training items. But rating a test item for acceptability only requires that groups of elements be characterized as more or less like the training groups. Likewise pattern completion requires no explicit hypothesis; what it requires is performance that is, to the extent possible, compatible with both the given elements and the regularities in the training set. For both behaviors, the learner takes a test item and attempts fit it into what is known of the pattern.

Even though the traditional characterization of induction does not precisely fit pattern learning, it is still the case that the learner is often faced with items which were not in the training set. The generalization exhibited by a model is a function of the training set, the form in which pattern elements and group positions are represented in the model, and the computational architecture and learning algorithm. We can distinguish between purely statistical learning algorithms and symbolic rule-learning algorithms.

Purely statistical learning algorithms keep track of co-occurrences of pairs of elements or element features, or co-occurrences of such pairs. That is, it is oriented toward content-specific correlational learning. A statistical algorithm may also keep track of what tends not to occur. Such a system will tend not only to be attracted to states which resemble those it has seen but also to be repelled by states which differ from what it has seen. Note that this latter capacity represents a potential solution to the problem of the lack of negative evidence, familiar from work on language acquisition (Marcus, 1993). Negative evidence often seems to be required to constrain the learner's hypotheses about what counts as acceptable, but we are assuming that negative evidence is not available during pattern learning.

Symbolic rule-learning algorithms (Marcus, 2000) differ from statistical learning algorithms in their explicit mechanism for abstracting away from data by replacing tokens with variables. Rules for patterns are among the simplest examples. Given training items in which groups consist of three elements and the first two elements of each group always belong to the same category, a rule-learning algorithm could form a rule characterizing the data of the form ($x\ x\ y$), where x and y match any element in the element domain. That is, it could learn relational correlations. This is an instance of what Thornton (2000) calls "relational learning", contrasting with the non-relational learning that would suffice for content-specific correlations. In order to work, rule-learning algorithms require a specification of what variability in the training data is required before an element is replaced by a variable in a rule and some means of determining what the element domain is (syllables, tones, polygons, etc.).

Pattern Learning in Infants

Now let's reconsider the experiments of Saffran et al. (1996) and Marcus et al. (1999) in terms of patterns. In the Saffran et al. experiments, infants learn about how certain specific syllables (actually syllable types) tend to be followed or preceded by certain other specific syllables. That is, this is an example of the learning of content-specific relational correlations. Although it has not been demonstrated in the experiments, the subjects may also learn content-specific higher-order

correlations. For example, if the syllable sequence *pidaku* is one of the recurring words in a pattern, the subject may learn that when *da* follows *pi*, it tends to precede *ku*.

In the Marcus et al. experiments, the infants are presented with groups of syllables, segmentable on the basis of both the first and the fourth of the four grouping characteristics described above: the intervals between groups are greater than those between elements within group, and each group embodies the same sort of regularity. It is this within-group regularity that is of interest. In our terms, the infants learn within-group relational correlations; they learn about the sameness or the difference between group elements in particular positions. For example, the infants in the (*x x y*) condition learn about the sameness relation between the first two elements within each group and the difference relation between the second and third elements (and possibly the first and third elements) within each group. Though it has not been demonstrated in these experiments, the infants may also learn higher-order correlations within the groups. For example, the subjects in the (*x x y*) conditions may learn that sameness between the first two elements co-occurs with difference between the second two elements in a group.

In sum, infants apparently have the capacity to learn both content-specific and relational correlations in patterns. But what sort of learning mechanism are they equipped with that allows them to do this? And does the relational learning case require something more sophisticated than statistical learning? In the next section we attempt to answer these questions in the context of a neural network model of correlational learning.

Pattern Learning in Neural Networks

What does it take to model the learning of patterns? In this section we describe features of the Playpen model of relational learning (Gasser, Colunga, & Smith, 2001) that are relevant for this task, specifically what is required to model the experiments of Saffran et al. and Marcus et al. In brief, the architecture is an attractor neural network with separate element units for each position in a pattern group and (for the Marcus et al. experiments) a set of hidden units representing primitive relations of sameness or difference on element dimensions in different group positions. The network is trained using a variant of Hebbian learning that punishes states which are not compatible with the pattern as well as rewarding states which are.

Architecture and Learning Algorithm

Our major claim is that pattern learning is correlational, and we start with the simplest architecture for correlational learning, an attractor neural network based on the continuous Hopfield model (Hopfield, 1984). A attractor network also has a natural way of implementing pattern completion and of yielding a measure of pattern goodness.

Units are joined by symmetric connections, and the network settles as unclamped units update their activations according to the familiar interactive activation update rule (McClelland & Rumelhart, 1981). A positive input moves a unit's activation towards its maximum, a negative input moves the activation towards its minimum, and a constant decay moves the activation towards a resting value. The units in the networks we describe below had a minimum activation of –0.2, a resting activation of 0.0, a maximum activation of 1.0, and a decay rate of 0.05.

There are two sorts of units, **input-output units** representing pattern element features or categories, and **relation units**, representing correlations between pattern element features or categories. Each relation unit receives input from and sends output to two input-output units. These connections, which share a single weight, are multiplicative; that is, the input into the relation unit is a function of the *product* of the weighted activations of the input-output units rather than the sum. Thus, in the absence of other input, a relation unit comes on only when it receives activation from both of its associated input-output units. Relation units may also be connected to one another by ordinary additive connections.

Each relation unit associates two values on an element dimension, represented by the two input-output units that it is connected to. Those associating the same value on the dimension (but in different pattern positions), **sameness units**, are distinguished from those associating different values on the dimension, **difference units**, by their pattern of connections. Sameness and difference are illustrated in Figure 2 below.

The learning algorithm is an unsupervised adaptation of contrastive Hebbian learning (Movellan, 1990), which augments simple correlational Hebbian learning with an anti-Hebbian learning phase that has the effect of punishing spurious states of the network that are not compatible with the training set. Specifically, during the positive phase of learning, a pattern is clamped on the input-output units, the hidden (relation) units are allowed to settle, and ordinary Hebbian learning is performed. That is, the weights are updated in proportion to the product of the activations of the connected units. (For the special connections joining input-output units and relation units, it is the product of the activations of all three units that is involved.) Next, in the negative phase of learning, the input-output units are unclamped, a small amount of noise is added to the activations of the units, and the network is allowed to settle again. Now anti-Hebbian learning is performed: the weights are updated in proportion to the *negative* of the product of the activations of the connected units. Note that if a training pattern behaves as an attractor in the network, during the negative phase the network should return to the state it was in following the positive phase (overcoming the injected noise). When this happens, the weight changes in the positive and negative phases for this pattern will cancel each other out. In other words, once the weight changes have stabilized, the network exhibits the desired behaviour.

Representation of Elements and Groups

As noted in the first section of this paper, pattern elements may be represented in either a local or a distributed fashion. In the local case, a single unit is assigned to each element type. This is illustrated in Figure 1a below. While this has the possibly desirable effect of treating elements as computational units in and of themselves, it leaves the system with no representation of inter-element similarity.

Distributed representations assign distinct sets of units to each element dimension, and each element is represented by the activation of multiple units. One possibility, then, is to turn on a single unit on each dimension for a given element. For example, say we are representing vowels. Vowels can be distinguished on several dimensions, the most important of which are (roughly) the relative height and backness of the nearest approach of the articulators in the oral cavity. With separate sets of units for each of these two dimensions, a vowel could take on the form of a pattern in which two units are activated and the others are not. Such a possibility is illustrated in Figure 1b, in which a single unit is assigned to each of five possible values along each dimension. For the vowel [o], two units, representing medium height and extreme backness, are strongly activated, and the other units are off or inhibited. While this style of representation allows the system to generalize on the basis of inter-element similarity, it fails to represent directly the ordering of values within dimensions. Instead we will assume that each element activates the units associated with a particular dimension with an on-centre-off-surround pattern. That is, we are really assuming that each of the element feature units has a Gaussian receptive field over the dimension and that neighbouring units have overlapping receptive fields. It is well-known that neural networks exposed to simple patterns such as activated pairs of adjacent input units can learn the weights for such receptive fields (von der Malsburg, 1973). Figure 1c illustrates this possibility for one vowel. As in Figure 1b, a single unit is assigned to each of five values on the two vowel dimensions. For the vowel [o], two units, representing medium height and extreme backness, are again strongly activated, but now the neighbours or these units are also weakly activated.

Figure 1 about here

For the different positions within pattern groups, there are two possibilities. Either the elements in different positions are represented across the same set of units, or separate units are allotted to each position. The latter alternative requires multiple units representing the same element categories or features. For example, in a network which processes sequences of syllables, there would be multiple copies of the syllable dimensions, one for each of the positions that is distinguished. While this may appear expensive, it is quite similar to a solution already instantiated in natural vision systems with multiple position-specific edge and motion detectors. This alternative also has the advantage that it avoids the binding problem (Shastri & Ajjanagadde, 1993). With distributed representations of multiple elements, this is the problem of keeping track of which activated unit goes with which object. For example, with multiple geometric figures and activated units representing CIRCLE, SQUARE, GREEN, and RED, how do we distinguish RED CIRCLE and GREEN SQUARE from GREEN CIRCLE and red square? Or with two syllables and activated units representing FRONT, BACK, HIGH, MID on the two vowel dimensions of backness and height, how do we distinguish syllables with vowels [i] (FRONT, HIGH) and [o] (BACK, MID) from syllables with vowels [e] (FRONT, MID) and [u] (BACK, HIGH)? The problem is solved with feature units specific to particular spatial or temporal positions, for example, CIRCLE-ON-THE-LEFT, GREEN-ON-THE-RIGHT, VOWEL1:HIGH, VOWEL2:MID. That is, units associated with features of elements in a particular position are in a sense grouped together in the network. Multiple copies of feature units also solve another problem: how multiple elements with the same value on a given dimension can be represented simultaneously.

In the model we propose, then, group positions (or relative positions within the sequence in the case of the simulation of Saffran et al.) take the form of distinct units. Relation units associate values on a given dimension across these positional groups of units. Figure 2 illustrates the pattern of connections. Two dimensions and two positions are shown. Within each positional copy of the dimension units, the units are arranged in the figure in the same order. That is, the first unit on the left represents the same value in the two rectangles in the first row. Sameness relation units (diamonds in the figure) associate the same value in different positions, whereas difference units (pentagons in the figure) associate different values in different positions. Other than this pattern of connectivity, the two types of relation units are identical. The dashed lines in the figure represent the multiplicative connections joining input-output units to relation units. Because the inputs to a relation unit are multiplied, it turns on (all else being equal) only when both of its input-output units are activated. This is illustrated for two sameness units and two difference unit in the figure.

Figure 2 about here

Pattern Behavior

In both sets of infant experiments, the subjects made implicit judgments of the acceptability of the test items. While there is nothing in the network that corresponds directly to looking time, the network has a natural means of indicating the deviance of a particular state in the extent to which it moves away from that state during settling. Specifically we test the acceptability of an item in much the same way as the network is trained. First the test sequence is clamped on the input-output units, and the network is allowed to settle. The activations of all of the units are recorded at this point. Next the input-output units are unclamped, and the network is allowed to settle again (without the injection of noise as during training). We now measure the Euclidian distance

between the network's state following the second settling and its state following the first settling. If the test sequence is familiar, the network should remain in roughly the same state, and the distance will be small. If the sequence is unfamiliar, the network should move toward a more familiar state, and the distance will be large.

Predictions

Simulating the experiments of Saffran et al. should be straightforward in a network of the type we have described since the differences detected by the infants are purely correlational. In fact the network for simulating these experiments should not require hidden (relation) units because there are no higher-order correlations involved.

The major claim of this paper is that it is also possible to simulate the learning of relational correlations, as in the experiments of Marcus et al., using simple correlational learning. What enables this in our model is distributed input representations based on overlapping receptive fields and relation units permitting higher-order relational correlations. While we can describe the patterns in these experiments with rules containing variables which are insensitive to content, the model, because it still operates by correlational learning, will always be sensitive to content. The upshot is that the network should "prefer" sequences to the extent that they resemble the training sequences. The most acceptable sequences should be those which are identical to the ones the network was trained on. At the other extreme should be sequences consisting of novel syllables which violate the training rule. In between would be sequences obeying the rule and consisting of novel syllables. Among these, acceptability should vary with the similarity of the component syllables to the training sequences. In the original experiments this range of possibilities is not tested; the test sequences consist only of very unfamiliar syllables either following or violating the training rule. We include more test sequences in our simulations to verify this aspect of the model's behaviour.

Simulations

Experiments of Saffran et al.

Architecture

In the experiments of Saffran et al. (1996), the element sequences were not pre-segmented for the subjects; the interval between adjacent elements within groups was the same as the interval between adjacent elements spanning group boundaries. Thus rather than assign separate input layers to group positions, we assigned separate layers to *relative* positions. We used four of these positional layers, one more than the length of the groups in the patterns.

Direct connections between the pattern units were sufficient to learn the correlations in the training sequence; there was no need for a hidden layer of relation units. The connection weights were initialized at 0.0. The syllables were represented in a local fashion; we discovered that networks given distributed syllable representations alone failed to make the distinction that the subjects made. The failure of distributed representations for this task is interesting since, as we explain below, distributed representations are required to simulate the Marcus et al. experiments with the model. It may be that infants have access to both syllable categories, such as *bi* and *ku*, and syllable features, such as the voicing of the onset consonant or the height of the vowel, and they capitalise on whichever captures the correlations in the data.

The architecture of the network is shown in Figure 3. Each input-output position layer contains twelve units, one for each syllable category; four of these are shown in the figure. All pairs of units in different position layers are joined by trainable connections. The four position layers represent a window of four consecutive syllables which the training sequence passes over as it is presented. That is, if the first six syllables in the sequence are *pa bi ku ti bu do*, the syllables

appearing on the four layers for the first three training trials are *pa bi ku ti*, *bi ku ti bu*, *ku ti bu do*. The first of these is shown in the figure.

Figure 3 about here

Training

As in the original experiments, the network was presented with 180 syllables in all. The three-syllable groups ("words") appeared together with equal frequency, and no group appeared twice in succession. Because each group consisted of unique syllables, this meant that the transition probability between any two syllables within words was 1, and the transition probability between the last syllable of a group and the first syllable of another group was 1/3. We trained three separate networks. Their behaviour was virtually identical since the only sources of differences were the order of the training trials and the order in which units were updated during settling. Not surprisingly, we also found that the negative phase of contrastive Hebbian learning is not necessary for this network with no hidden units.[2]

Testing and Results

The task in these experiments is not generalization to novel sequences; it is to make a distinction between more and less likely sequences. Sequences constituting groups or subsequences within group in the pattern should be more acceptable in some sense than sequences spanning groups. We used the method described above to measure acceptability (because there was no hidden layer, no initial settling was required). To simulate the presentation of three-syllable test sequences, we presented each sequence to the first three or the last three of the four layers of the trained network, starting the units in the other layer at values close to their resting activations. We then let the network settle and measured the change in activation, that is, the Euclidian distance between the states before and after settling.

The mean change in activations following the presentation of three-syllables sequences constituting words was 0.55. The mean change following three-syllable sequences spanning word boundaries was 0.88 ($t(27) = 159$, $p < 0.0001$). The network has clearly learned to distinguish within-group from between-group transitions.

Discussion

In sum, a simple Hopfield network replicates the results of the experiments of Saffran et al. involving content-specific correlations in a pattern learning task. These results are not particularly surprising — after all, this sort of statistical learning is what neural networks excel at — we include it mainly to clarify what the architecture requires beyond this to learn the relational correlations in the experiments of Marcus et al.

As noted in the discussion of pattern behaviour, people are capable not only of distinguishing familiar from unfamiliar patterns. They can also produce patterns that agree with the regularities they have learned. Pattern completion is a natural process in attractor neural networks such as these, and we tested the networks for this simulation to see whether they could generate portions of pattern groups given others. To do this, we clamp some of the input-output units and then let the unclamped units settle. The network should fill in the unclamped dimensions with values compatible with the clamped units and with the regularity in the training patterns.

---

[2] Contrastive Hebbian learning (Movellan, 1990) augments the usual Hebbian learning used to train simple attractor neural networks without hidden units (Hopfield, 1984) with the anti-Hebbian learning of the negative phase of training. The hidden units in a network represent re-encodings of the input patterns which permit more complex mappings to be learned. However, the initial random weights to the hidden units may lead to spurious attractors. The negative phase of training is designed to eliminate these spurious attractors by effectively penalizing states which do not yield appropriate input-output mappings.

We tested each of the four word sequences by clamping units for the first two syllables in either the first and second or second and third group layers and leaving the other layers unclamped. We then allowed the unclamped units to settle. At issue was whether the layer following the first two syllables would settle to the third syllable in the sequence, that is, whether the unit for that syllable would be more highly activated than the others in that layer. In all cases, this was what we found.

Experiments of Marcus et al.

Architecture

In the experiments of Marcus et al. (1999), element groups were clearly delineated for the subjects by intervals longer than the inter-element intervals within groups. Thus we assumed a pre-processing stage in which the group positions were assigned to distinct sets of element units. That is, there were three layers of units for the representation of syllables.

Local representations of syllables would not suffice to simulate these experiments because the network must generalize on the basis of the similarity between the syllables. We represented syllables using a minimal set of dimensions, two for consonants and two for vowels. The consonant dimensions were **sonority**, roughly the extent to which a consonant is vowel-like, and **place of articulation**, the position in the vocal tract of the narrowest contact between articulators. The vowel dimensions were **height** and **backness** of the narrowest approach between articulators. All four of these dimensions are accepted ways of characterising phonetic segments. Normally more dimensions would be required, but these four suffice to distinguish the phones used in the third experiment of Marcus et al.: *d*, *j*, *l*, *w*, *e*, *i*. Each of the dimensions was represented by five units with overlapping receptive fields. Thus the presentation of an element with a given value on a dimension always activated one unit on that dimension strongly, activated its two neighbouring units (or one neighbouring unit if it was at the end of the scale) weakly, and inhibited the other units on the dimension.[3]

We discovered, to our surprise, that it was possible to train a network with no hidden units to distinguish the syllable sequences distinguished by the infants in the experiments, but this required more presentations of the patterns and the difference in the network's familiarity with the two kinds of sequences was small, not a convincing simulation of the differences in the infants' looking times.

The addition of relation units allows the network to make use of correlations of particular co-occurrences along dimensions. The most obvious of these is the non-occurrence of both same and different values on a dimension. For example, given two successive syllables, the heights of their vowels were either the same or different. Less obvious is the tendency for sameness on one dimension to correlate with sameness on the other dimensions for a given pair of positions. This was true for the stimuli because pairs of syllables were either identical, or they were significantly different. However, it was only a tendency; sometimes the different syllables shared a vowel or a consonant (*ji*, *wi*; *ji*, *je*).

For every pair of units on corresponding dimensions in different positions, there was an associated relation unit. Thus there was a sameness relation unit for a sonority of 0 in group position 1 and a sonority of 0 in group position 2; this unit tended to be activated when the consonants of the first two syllables in a group both had sonority 0 (that is, when both were voiceless stops). Likewise there was a difference relation unit for a sonority of 0 in group position 1 and a sonority of 2 in group position 2. The relation units were connected to their associated input-output units with small constant weights, smaller for the difference units because there were more of these. These weights were modified during learning.

[3] In the case of *w*, there were two places of articulation, labial (0) and velar (3). Both of the corresponding units received relatively high, though not maximum, activation.

In addition to these connections there were three others kinds. For a given dimension and a given pair of group positions, difference and sameness units inhibited one another, reflecting the first sort of correlation discussed above. These small negative weights were not modified during learning since they presumably represent basic knowledge about the incompatibility of sameness and difference relations within a dimension which would have been learned before the experiment.

There were also connections among all of the sameness units on different dimensions for each pair of group positions. These trainable connections had initial weights of 0.0 and would permit the network to learn the second sort of correlation discussed above: correlations between sameness relations on different dimensions for a given pair of group positions, for example, the tendency for sameness in vowel height to correlate with sameness in vowel backness.

Finally there were connections permitting the learning of correlations between co-occurrences of values for different pairs of group position, for example, the tendency for sameness on vowel height between positions one and two to correlate with sameness in vowel height between positions two and three. There were actually few such correlations in the training sequences, and the inclusion of these connections made little difference in performance, but they were included for the sake of completeness. Figure 4 shows the architecture of the network for simulating the experiments of Marcus et al. The input-output units (white rectangles), representing position-specific syllable features, include four dimensional layers for each group position. For every pair of group positions there is a set of four dimensional layers for sameness relation units (dark grey rectangles) and four dimensional layers for difference relation units (light grey rectangles). The dashed lines summarize the multiplicative connections joining pairs of input-output units with each relation unit. Connections between relation units are not shown.

Figure 4 about here

Figure 5 shows detail for the connections between the input-output units for a single dimension in two different positions and the corresponding sameness relation units. Two different input patterns are shown to illustrate how the relation units are activated in response to them. Difference units behave in the same fashion (see also Figure 2). However, there are more of them for each pair of position-specific dimensions, one for each combination of a value on one input-layer with a different value on the other input-layer, that is, 20 units for our network with 5 units per input-output dimension layer.

Figure 5 about here

Training

Each network was trained on sequences obeying a particular "rule". Because the network's representations of different positions within a sequence were identical, there was no reason to train separate networks on the three rules. We will assume that the training rule is (*x x y*). The 16 training syllable sequences were those used in Experiments 2 and 3 of Marcus et al.: *le le di*, *le le we*, *le le li*, *le le je*, *de de di*, *de de we*, *de de li*, *de de je*, *ji ji di*, *ji ji we*, *ji ji li*, *ji ji je*, *wi wi di*, *wi wi we*, *wi wi li*, *wi wi je*.

As in the original experiments, the networks were presented with each training sequence three times during training. Contrastive Hebbian learning was used to train the network. A training sequence was clamped over the input-output units, the hidden units were allowed to settle, and Hebbian weight updates were accumulated. Then the input-output units were unclamped, a small amount of noise was added to the units' activations, and all units were allowed to settle. Finally, anti-Hebbian weight updates were accumulated.

We trained three separate networks. None of the initial weights were random, and the minor variations in network performance we observed were due to the random order of training sequences and of units to update during network settling.

Testing and Results

We tested the network on the test sequences from the original experiments, made up of the syllables *ba*, *po*, *ko*, and *ga*, as well as on a set of other sequences consisting of syllables either from the training sequences or resembling those in the training sequences.

We modelled the infants' differential familiarity as described above. A test sequence was clamped on the input-output units, the hidden units were allowed to settle, and the activations of all units were recorded. Next the input-output units were unclamped, and all of the units were permitted to settle again. The state of the network was then compared to the saved state.

Figure 6 shows results for four categories of test sequences: training; grammatical, partially familiar; grammatical unfamiliar; ungrammatical unfamiliar. "Grammatical" means following the training rule. "Partially familiar" refers to sequences containing syllables that overlap significantly with the training syllables. For example, *jo* resembles the training syllables *ji* and *je*. Each "partially familiar" syllable combined a training consonant with a test vowel or a test consonant with a training vowel. Clearly the network exhibits the basic effect found in the experiment. Grammatical patterns consisting of unfamiliar syllables are preferred over ungrammatical patterns consisting of those same unfamiliar syllables ($t(16) = 10.8$, $p < 0.001$).. In addition, the network makes several predictions about the extent to which subjects will generalise to novel sequences.

Figure 6 about here

Discussion

How is it that the network appears to have rule-governed behaviour when it is only learning correlations? The answer lies in the overlap between the different syllable representations and in the form that sameness takes in the network. Marcus et al. strove to define a set of training and test syllables that would not overlap at all. But *w*, one of the training consonants, involves both bilabial (0 on our place of articulation dimension) and velar (3) articulation, so training on syllables containing this consonant in the positions where the same syllables occur exposes the model to sameness on two points along the place of articulation dimension. These happen to be the same two points which apply to the consonants in the dissimilar test syllables, that is, *p*, *b*, *k*, and *g*. Likewise, both of the training vowels, *i* and *e*, have medium height (2 on our height dimension), like one of the test vowels, *o*. The receptive fields of the input-output units, each encompassing more than one value on the its dimension, also lead in general to more commonality between syllables. It is true that there is no overlap at all between the training and test syllables along the dimension of vowel backness. But given the distributed representations of syllables, there is plenty of basis for generalisation to the test syllables within the other three dimensions.

Thus the network has no abstract notion of sameness. Sameness between two sequence group positions is nothing more than a strong association (implemented in the form of a relation unit) between every pair of element features which match one another. And in the absence of exposure to examples spanning the range of values on each dimension, we can expect the learner to respond with varying commitments to sameness.

In other words, content still matters. Rather than a pristine variable which matches anything in its domain, we have a whole array of units, some readily activated, others less likely to respond. As we have seen, the model predicts different responses for sequences depending on their similarity to training sequences. We are currently performing an experiment using visual patterns and adult subjects to test this prediction. 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 sequences, as predicted by our model.

But in a sense content matters for the rule-based account too. Variables are normally defined over some domain (syllables, two-dimensional figures, etc.). They match all elements within the domain equally well but fail to match elements outside the domain. In the model we are proposing, there may be no such clean lines. Just as we expect elements within the domain to match the "rule" to different degrees, there is no reason not to expect elements outside the domain to match to some extent as well. What matters is how much they overlap with the familiar elements. In fact without variables there may be no more reason to posit element domains.

There is a further sense in which symbolic models are usually all-of-none. The learner is presented with a set of training items and for some time can generalise only on the basis of similarity. Then at some point the learner "gets" the rule. But at what point? How much training and what quality of training are required for this to happen? A statistical model like the one we are proposing requires no such threshold; learning is a continuous phenomenon. Again the predictions differ; the symbolic model predicts a significant discontinuity in performance.

As we have already seen, pattern knowledge in older children and adults goes beyond the passive ability to discriminate between sequences on the basis of how well they agree with the pattern. People are also able to produce patterns of the type used in the experiments of Marcus et al. A good example of this is the ability of speakers of languages making use of productive morphological reduplication (Moravcsik, 1978) to generate the reduplicated forms of stems. In reduplication one portion of a stem is copied somewhere in the word, as in Tagalog *bibili* 'will buy' from *bili* 'buy'. As noted above, this may take the form of pattern completion, the production of some group elements given others.

As in the Saffran et al. simulations, we tested this ability in the networks using pattern completion. For the networks used in simulating the Marcus et al. experiments, the results were more negative. Having trained a network on the (*x x y*) rule, we clamped the *y* syllable and one of the *x* syllables and allowed the units representing the other syllable to settle. For the training sequences, the network readily filled in the appropriate syllable. For the most unfamiliar syllables, those used in the test sequences in the experiments of Marcus et al., the network correctly filled in only one or two of the four syllables at best. For the intermediate case of partially familiar syllables, the correct syllable was filled in only 1/3 of the time; more often only two or three of the four features were generated correctly.

These results do not indicate that the model cannot learn to produce patterns involving relational correlations. Pattern completion is a much more demanding task than simply ranking one sequence as more compatible with the pattern than another. The model would clearly require training on a greater variety of elements to perform this task. We do not know what it takes for a language learner to attain the ability to perform reduplication. This is an area we hope to investigate in artificial language learning experiments.

In addition to these large-scale predictions, the model makes a number of detailed predictions. For example, in the model of the Marcus et al. experiments, performance varied significantly within a category of test sequence. This was especially true for stimuli containing the partially familiar syllables, those composed of combinations of familiar and unfamiliar phones. For some reason, the sequence *la la wo* was not readily accepted by the network; in fact for these syllables all three network instantiations actually preferred the ungrammatical sequences (*la wo wo*, *la wo la*). In general the network was less willing to accept novel vowels than novel consonants. This was probably due to the relatively small variation among the training vowels (two values for one dimension, one value for the other).

Conclusions

We began with patterns and pattern learning and described how very young children are already capable of learning the two kinds of correlations in patterns. These capacities will be fundamental in later learning, of language in particular. They will play a role in segmentation generally and in phonological development specifically. Understanding the mechanisms behind these capacities is fundamental to understanding language and cognition itself.

In struggling for this understanding, we should be aware that a given set of regularities has multiple characterisations and that the characterisation that is most obvious to us as observers of the learning process, rather than as learners, may not be the one that learners pick up on. One advantage of simple statistical learning devices that make use of distributed representations is that they can sometimes reveal these less obvious sorts of characterisations to us. Or at least thinking in terms of these models and what they can do may liberate us from thinking more abstractly than we need to.

We have shown that there is such an alternate characterisation of the artificial stimuli used in the experiments of Marcus et al. (1999), one that makes no reference to variables or formal rules. Given this sort of characterisation, building a statistical model that learns the regularities is straightforward. Other connectionist modelers have also discovered alternative ways of viewing this seemingly symbolic task. As we investigate and re-investigate the more complex tasks that make up language and higher cognition, we should keep this in mind..

References

Christiansen, M.H., Conway, C.M. & Curtin, S. (2000). A connectionist single-mechanism account of rule-like behavior in infancy. *Annual Conference of the Cognitive Science Society*, *22*, 83-88.

Gasser, M. & Colunga, E. (2000). Babies, variables, and relational correlations. *Annual Conference of the Cognitive Science Society*, *22*, 160-165.

Gasser, M., Colunga, E., & Smith, L. B. (2001). Developing relations. In E. van der Zee & U. Nikanne (Eds.), *Cognitive interfaces: Constraints on linking cognitive information*. Oxford, UK: Oxford University Press.

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.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition* , *46***,** 53-85.

Marcus, G. F. (2001). *The algebraic mind: integrating connectionism and cognitive science.* Cambridge, MA, USA: MIT Press.

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

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375-407.

Moravcsik, Edith A. 1978. Reduplicative constructions. In J. H. Greenberg (Ed.), *Universals of human language*, *vol. 3*. *Word structure*. Stanford, CA, USA: Stanford University Press.

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, J. L. Elman, & T. J. Sejnowski (Eds.), *Proceedings of the 1990 Connectionist Models Summer School*. San Mateo, CA, USA: Morgan Kaufmann.

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

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

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70,* 27-52.

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.

Shultz, T. R., & Bale, A. C. (2000). Infant familiarization to artificial sentences: rule-like behavior without explicit rules and variables. *Annual Conference of the Cognitive Science Society*, *22*, 459-463.

Seidenberg, M. S., Elman, J. L., Negishi, M., Eimas, P. D., & Marcus, G. F. (1999). Do infants learn grammar with algebra or statistics? *Science*, *284*, 433.

Thornton, C. (2000). *Truth from trash*. Cambridge, MA, USA: MIT Press.

von der Malsburg, C. (1973). Self-organization of orientation selective cells in the striate cortex. *Kybernetik*, *14*, 85-100.

Figure 1

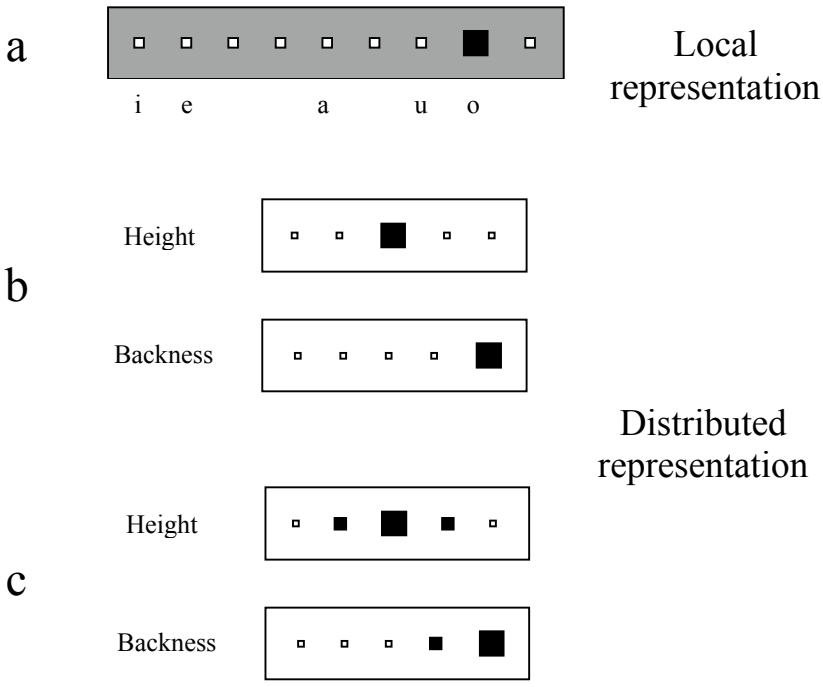

a
i e a u o
Local representation
b
Height
Backness
Distributed representation
c
Height
Backness

Figure 2

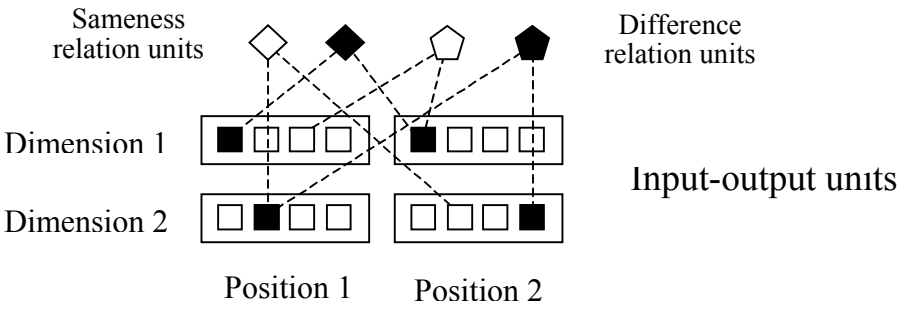

Sameness
relation units
Difference
relation units
Dimension 1
Dimension 2
Input-output units
Position 1
Position 2

Figure 3

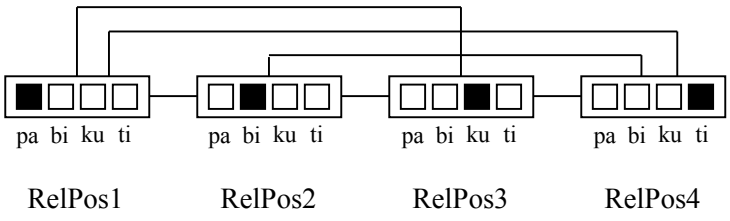

pa bi ku ti
pa bi ku ti
pa bi ku ti
pa bi ku ti
RelPos1
RelPos2
RelPos3
RelPos4

Figure 4

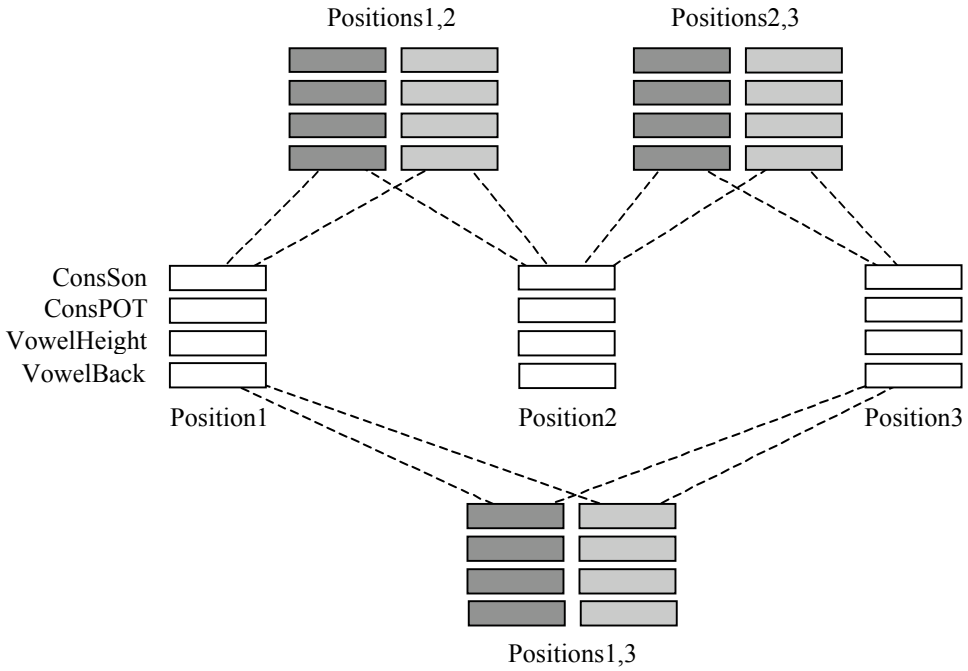
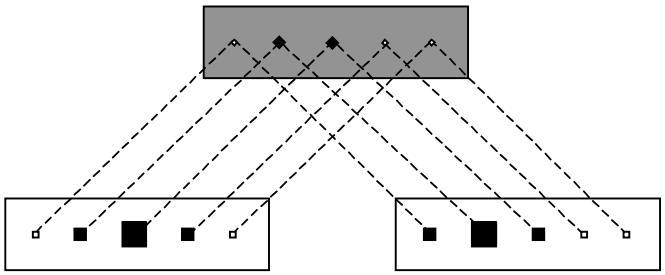

Positions1,2
Positions2,3
ConsSon
ConsPOT
VowelHeight
VowelBack
Position1
Position2
Position3
Positions1,3

Figure 5

a



b

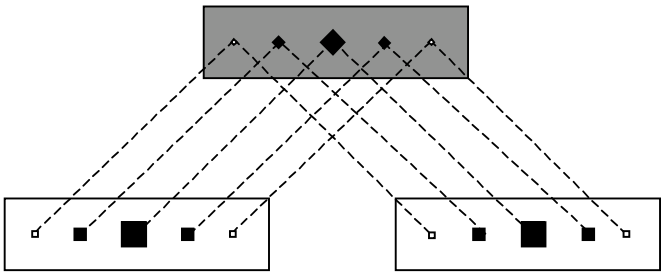


Figure 6

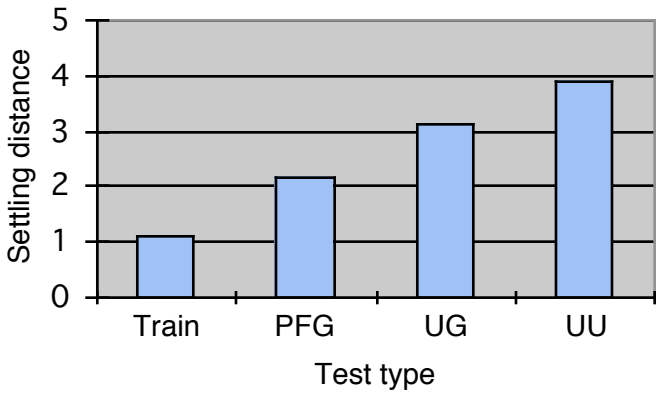


Figure Captions

Figure 1. Local (a) and distributed (b, c) representations of pattern elements, illustrated for vowels. Black squares represent positively activated units, white squares negatively activated units. The size of a square represents the magnitude of activation. For the distributed representations, there are two vowel dimensions, height and backness. Each vowel is characterized by a vowel on each of these dimensions. In b, the distributed representations are local within each dimension; in c, they are distributed within each dimension as well.

Figure 2. Basic architecture of the model. Input-output units represent element features (dimension values) or categories. Relation units associate input-output units along multiplicative connections (dotted lines). Separate units represent positions within pattern groups. Sameness relation units (diamonds) associate the same feature in different positions. Difference units (pentagons) associate different features in different positions. Activated units are indicated by filled figures..

Figure 3: Architecture of network for simulating experiment of Saffran et al. (1996). Relative position layers consist of input-output units. Units in different layers are joined by trainable connections, indicated by the lines in the figure. Within each position layer only four of the twelve units are shown. The activated units in the figure represent the subsequence *...pa bi ku ti* ...appearing somewhere within the input sequence of syllables.

Figure 4. Architecture of network for simulating experiments of Marcus et al. (1999). There are input-output units (white rectangles) for each syllable dimension in each of the three group positions. Each pair of positions has associated layers of sameness (black rectangles) and difference relation units (grey rectangles). Not shown are connections between relation units.

Figure 5. Detail of connectivity between input-output and relation unit layers. Shown are the input-output layers for one syllable dimension in two group positions and the associated layer of sameness relation units (diamonds). The response of the sameness units to two different on-centre-off-surround input patterns is shown, one in which the value on the dimension is different for the two positions (a) and one in which the value is the same (b).

Figure 6. Simulation of experiment of Marcus et al. (1999). Distance between networks states before and after unclamping of IO units. Test types: Training; Partially Familiar, Grammatical; Unfamiliar, Grammatical; Unfamiliar, Ungrammatical