LANGUAGE AND COGNITIVE PROCESSES, 2000, *15* (1), 87–127

Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant

Peter Ford Dominey
Institut des Sciences Cognitives, Bron, France

Franck Ramus
Laboratoire de Sciences Cognitives et Psycholinguistique, Paris, France

Well before their first birthday, babies can acquire knowledge of serial order relations (Saffran et al., 1996a), as well as knowledge of more abstract rule-based structural relations (Marcus et al., 1999) between neighbouring speech sounds within 2 minutes of exposure. These early learners can likewise acquire knowledge of rhythmic or temporal structure of a new language within 5–10 minutes of exposure (Nazzi et al., 1998). All three of these types of knowledge likely play invaluable roles in "bootstrapping" language acquisition. Two important open questions that remain include: What are the mechanisms that provide this rapid learning ability, and how do they depend on pre-exposure to the environment? Here we show that a neurophysiologically validated temporal recurrent network simulates babies' capabilities to learn serial order and rhythmic structure. Indeed the recurrent network is capable of representing serial and temporal structure with no pre-exposure, and through exposure these internal representations can become bound to behavioural responses. In contrast, babies' performance in extracting abstract structure can only be simulated by a modified version of the model. We thus demonstrate how innate representational capabilities for serial and temporal structure of language could arise from a common neural architecture, distinct from that required for the representation of abstract structure, and we provide a predictive testable model of at least these aspects of the initial computational state of the language learner.

Requests for reprints should be addressed to Peter F. Dominey, Institut des Sciences Cognitives, CNRS UPR 9075, 67, Blvd Pinel, 69675 BRON Cedex, France. Telephone: 04 37 91 12 66. Fax: 04 37 91 12 10. E-mail: dominey@isc.cnrs.fr

This research was supported by the GIS Sciences de la Cognition (Paris). FR was supported by a grant from the DGA. The authors wish to thank the anonymous reviewers for their extensive questions, clarifications and comments that significantly improved the quality of the final manuscript.

INTRODUCTION

An important part of language acquisition consists in extracting regularities directly from the speech signal. Indeed, many areas of language, including aspects of syntax and the lexicon, seem to be unlearnable without the help of external cues. It has thus been proposed that prosodic cues, and more generally phonological cues, are correlated with more abstract properties of language, and hence can help "bootstrap" their acquisition (Gleitman & Wanner, 1982; Christophe et al., 1997; Morgan, 1986; Morgan & Demuth, 1996). To be tenable, this view crucially requires researchers to determine the nature and the reliability of the cues that can be extracted directly from the speech signal, and to show that these cues are indeed used by infants in the acquisition process. Statistical analyses of speech corpora allow the discovery of potentially useful regularities (see for example, Cutler & Carter, 1987; Brent & Cartwright, 1996; Reddington, Chater, & Finch, 1998; Shi, Morgan, & Allopenna, 1998). Perceptual studies ensure that these regularities are indeed perceived by infants (see Jusczyk (1997) for a review). In these two respects, artificial neural network models can play a role. Given their general regularity detection capabilities, they can help discover new and correlated regularities in the input (Christiansen, Allen, & Seidenberg, 1998; Elman, 1990, 1991, 1993; Seidenberg, 1997). Such sensitivity should allow these models to simulate the regularity-extraction performance of infants in perceptual experiments. Depending on their neurophysiological realism, models that succeed stimulating certain capacities may contribute to an understanding of how these capacities are implemented in the neural tissue. In this paper, we present such a neural network model, that may contribute to the understanding of infants' sensitivity to serial, temporal and abstract regularities in language.

In behavioural sequences including speech production, music, skilled motor control like walking, dancing or typing etc., the temporal organisation of the sequence is of nearly the same importance as the serial order of events, and the two are quite often well correlated. In music, a melody is defined not only by the serial order of the notes, but also by their durations and the pauses between them. Similarly, such temporal information is crucial in speech, both for naturalness and comprehension, as robotic speech featured in science-fiction movies reminds us. Behavioural sequences can also be organised around more abstract structures or rules that permit the generation of new but "legal" instances, with syntax being a good example of such an abstract structure. In language, regularities in the serial structure concern the distribution of phonemes and syllables, and may help in finding word boundaries in fluent speech. Regularities in the temporal structure reflect global rhythm, emphasis,

lexical accent, etc., and regularities in the abstract structure reflect morphological and syntactic rules. All these types of regularities are likely exploited by the infant to help bootstrap language acquisition. Indeed, data from the study of language acquisition demonstrates sensitivity to the serial (Jusczyk, Luce, & Charles-Luce, 1994; Mandel, Jusczyk, & Kelmer Nelson, 1994; Morgan, 1994; Morgan & Saffran, 1995; Saffran, Aslin, & Newport, 1996a), temporal (Christophe et al., 1994; Hirsch-Pasek et al., 1996; Mandel et al., 1994; Nazzi, Bertoncini, & Mehler, 1998) and abstract (Marcus et al., 1999) structures of natural and artificial languages in human infants.

We have previously argued that serial and temporal or rhythmic structure can be naturally treated by a common mechanism in sensori-motor sequence learning (Dominey, 1998a, b), while abstract structure must be processed by a dissociated system (Dominey, 1997; Dominey et al., 1998). We now address the possibility that this dissociation holds within the domain of language acquisition in the baby as well. That is, for the baby, serial and temporal structure in language may also be treated by a common mechanism, while abstract structure is treated by a separate and dissociated mechanism. In order to approach this problem, we first review three studies that address, respectively, the sensitivity to serial, temporal and abstract structure in the baby. These studies were selected, in part, because they test infants performance on speech material that is not from the native language, thus probing the capabilities to form representations in real-time, independent of specific experience with the material to be represented.

Serial/distributional structure

In order to acquire a lexicon and the syntactic rules that apply over words, infants first need to extract words from fluent speech. Although word boundaries are not consistently marked in speech like in writing, there are partial cues that can help in this task. Phoneme sequences spanning across word boundaries are typically less frequent than those inside words (Harris, 1954, 1955). Moreover, there are phonotactic constraints over the consonant clusters that may occur word-internally, compared to those occurring across word boundaries. Computer simulations show that these and related serial ordering regularities are useful to solve the word segmentation problem (Brent & Cartwright, 1996; Christiansen et al., 1998), and perceptual studies have shown that before the end of their first year, infants are sensitive to the phonotactic constraints of their maternal language (Friederici & Wessels, 1993; Jusczyk, Charles-Luce, & Luce, 1994). Furthermore, Saffran, Newport, & Aslin (1996b) have given

evidence for an additional cue, the transitional probabilities between successive syllables. These observations indicate that before their first birthday, infants are sensitive to serial ordering regularities in their native language. It is not clear, however, how these infants would fare in extracting serial order regularities in real-time from novel stimuli not in their native language. Saffran et al. (1996a) have addressed this issue by looking at the ability of babies to exploit serial ordering regularities in an artificial language made up of nonsense words, thus eliminating the possibility of previous exposure to these words. In the first of two experiments, 8-month-old babies were exposed to 2 minutes of a continuous speech stream made up of four three-syllable nonsense words repeated in a random order, for a total of 180 words during the 2 minute period. After this training period, the babies were then exposed to a testing period that involved random presentation of 4 words, two that were words used during the preceding training, and two that were new, made up of novel combinations of the previously presented syllables (see Table 1).

During the test phase, the infants demonstrated a significant selective sensitivity to the test stimuli, with significantly longer listening times for the novel non-words. The idea is that syllable pairs within the words of the training corpus have high transition probabilities. In sharp contrast, the syllable pairs that occurred in the non-words had never occurred as pairs during the training and thus had transition probability of zero. The babies can thus make this discrimination in the test phase, recognising words that have the same serial order of syllables as those seen in training.

The second experiment tested the more difficult discrimination between words that occurred during training vs. "part-words" made up of syllable strings that spanned word boundaries during the training. Thus, in this experiment the part-words had actually occurred during the training, but with reduced frequency with respect to the words. Again, two minutes of exposure was sufficient to allow 8-month-old babies to successfully perform this discrimination between high frequency words and lower frequency part-words during the subsequent testing phase. This indicates that the babies are not only sensitive to the serial order of sound-strings in the training stream, but also to the distributional frequency of occurrence of these sound-strings.

Temporal/rhythmic structure

While the serial or distributional organisation of speech sounds is clearly an important source of information to the language learner, it is equally clear that it is not the only one. To consider again the word segmentation problem, there are important temporal cues to word boundaries in addition to distributional cues. In English for instance, where the majority

of words have stress on the first syllable, word boundaries are likely to be found before the longest syllables (Cutler & Butterfield, 1990). More generally, the actual durations of phonemes and syllables and the pauses between them, together with other cues like pitch and energy, provide information concerning both word- and clause-boundaries (Cooper & Paccia-Cooper, 1980; Klatt, 1976; Nakatani & Shaffer, 1978). Again, infants seem to be sensitive to these cues as well, although the respective roles of duration, pitch and energy have not yet been clearly disentangled (Christophe et al., 1994; Hirsh-Pasek et al. 1987; Jusczyk & Aslin, 1995; Mandel, Jusczyk, & Kelmer Nelson, 1994; Morgan, 1996). Furthermore, different languages have different global rhythmic or temporal structure (Abercrombie, 1967; Ladefoged, 1975; Pike, 1945): they are usually classified as stress-timed, syllable-timed or mora-timed. The early ability to classify one's native language into one of these three rhythm classes is also thought of as a potential bootstrap, possibly cueing more abstract phonological properties, like syllable structure (Mehler et al., 1996; Ramus, Nespor, & Mehler, in press). Already within the first days after birth, human infants are capable of discriminating between unfamiliar languages from different rhythm classes based on prosodic information (Mehler et al., 1988; Nazzi et al., 1998). In these experiments sentences that have been low-pass filtered to preserve only the prosody are presented out loud to the infants, and the behavioural measure is the rate of sucking on a pacifier. After habituation to sentences in one rhythm class, discrimination is observed as an increase in sucking rate when sentences in a different rhythm class are presented in the test phase. In contrast, no change is observed in the control groups when sentences from the same rhythm class (but different speakers) are presented in the test phase. In Nazzi et al. (1998) infants discriminate between stress-timed English and mora-timed Japanese (Experiment 1), but fail to discriminate between stress-timed English and stress-timed Dutch (Experiment 2). In Experiment 3, infants heard different combinations of sentences from English, Dutch (stress-timed), Spanish and Italian (syllable-timed). Discrimination was observed only when a mixture of English and Dutch sentences was contrasted with a mixture of Spanish and Italian sentences. Only in this case were sentences from one rhythm class contrasted with sentences from a different rhythm class. These results demonstrate a general sensitivity to the rhythmic structure of language in babies, that can likely be exploited to permit the extraction of supplementary linguistic regularities.

Abstract structure

The previous two studies indicate that babies can extract statistical regularities both in terms of the serial ordering, and the timing

regularities within sound sequences. A recent study by Marcus et al. (1999) has demonstrated that 7-month-old infants can also extract abstract regularities in sentences of an artificial language and can transfer this knowledge to entirely new sentences made of "words" the infant has never heard before. Using a familiarisation preference procedure as adapted by Saffran et al. (1996a), two groups of infants were exposed to either an ABA or an ABB grammar condition. In the ABA condition, the infants were familiarised with a two-minute synthesised speech sample of 16 sentences (3 repetitions of the 16 sentences in a random order) that followed the ABA grammar such as "ga ti ga" and "la ni la" (see Table 3), and sentences like "ga ti ti" in the ABB condition. In the transfer test phase, infants were exposed to sentences made up of entirely new words such as "we fe we" or "we fe fe". Half of the test trials were consistent with the training grammar, and half were from the other grammar not used in training. This transfer test measured whether abstract knowledge acquired in training would transfer to new sentences (constructed from new words) consistent with the learned grammar. Fifteen of sixteen infants demonstrated a looking preference for inconsistent sentences, indicating that they had indeed acquired and transferred knowledge of the grammar.

Two further experiments were performed to resolve control issues. Experiment 2 was used to eliminate the possibility that infants were relying on unintended phonetic regularities common to training and testing items. Thus, the grammars were the same as those used in Experiment 1, after elimination of any phonetic regularities linking the training and testing sets. The third experiment addressed the criticism that to distinguish ABA from ABB, babies might just be sensitive to immediate reduplication (i.e., XXY vs. XYX) in which case they would fail to distinguish ABB from AAB. Thus, Experiment 3 used the phonetically neutral words from Experiment 2, with grammars AAB and ABB (see Table 3). In both of these control experiments the original observation of learning and transfer of the abstract structure was maintained.

Thus, infants at seven months are capable of extracting an abstract grammar-like structure from a set of training sentences and transferring this information to new sentences that are made up of different words. This capability for rule extraction and the capability to extract statistical regularities appear to be useful, though not sufficient, components of a language acquisition capacity. Marcus et al. (1999) note that while standard sequence learning models such as the simple recurrent network (SRN) (Elman, 1990) can pick up on sequential regularities, they would fail in this kind of transfer task, since the knowledge acquired by these models is specific to the elements that are used to make up the sequence.

Dissociable systems

We have thus seen the rather impressive sensitivity in babies to the serial, temporal and abstract structure of events in language as revealed by Saffran et al. (1996a), Nazzi et al. (1998) and Marcus et al. (1999) respectively. We now attempt to address the following questions: What are the mechanisms underlying these specific forms of sensitivity to serial, temporal, and abstract and structure? Are there multiple mechanisms, or can all three of these functions be realised by a single shared mechanism? Additionally, we can ask, what pre-exposure to the linguistic environment is required for such a mechanism(s) to properly function? Answering these questions will provide an important step in establishing how phonological structure can be used to bootstrap the acquisition of morphological and syntactic structure.

We have previously demonstrated that a temporal recurrent network (TRN) based on the neuroanatomy of the primate frontostriatal system (Dominey, Arbib, & Joseph, 1995) can learn both the serial and the temporal structure of sensorimotor sequences (Dominey, 1998a,b), but that it fails to learn their abstract structure which requires architectural modifications (Dominey, 1997; Dominey et al., 1998). In this paper we first attempt to determine if the TRN model can demonstrate the same sensitivity to serial and temporal structure in language as that of babies by exposing it to the experimental conditions reported in Saffran et al. (1996a) and Nazzi et al. (1998). We then attempt to validate the claim that abstract learning as demonstrated by Marcus et al. (1999) can only be achieved by a system that has specific representational capabilities for abstract structure, elaborated here as an Abstract Recurrent Network (ARN). We start by providing a brief description of the initial TRN model, with an emphasis on the characteristics that distinguish it from related models.

THE TEMPORAL RECURRENT NETWORK (TRN)

A critical aspect of sequence learning is the ability to deal with ambiguous sequences such as DCABCABAC in which a given element (such as B) has different successors (DCAB is followed by C, while BCAB is followed by A). Resolving these ambiguities requires that the system can store some number of previous elements—a context—in order to resolve the ambiguity. The complex sequence learning capacity of recurrent networks has been demonstrated in a variety of settings where the encoding of previous states or events allows prediction of future events (Cleeremans & McClelland, 1991; Dominey, 1995; Elman, 1990). Recurrent connections allow information from previous states of the system to play back into the

current state, with the result that the current state of activation represents not only the current inputs but also on the context of previous inputs.

In a number of recurrent network studies, learning modifies the recurrent connections (Christiansen et al., 1998; Cleeremans & McClelland, 1991; Elman, 1990). If one is interested in treating the temporal structure accurately, e.g., preserving the different durations of prosodic events as in the Nazzi et al. (1998) study, one faces the interesting technical challenge of how to keep track of the role played by a given recurrent connection over multiple simulation cycles that correspond to a given duration (Pearlmutter, 1995; Pineda, 1989). A standard solution to this problem is to use a uniform temporal structure so that on each simulation time step or cycle, the next input and output sequence elements are processed, and variation in the durations of individual elements is lost (Christiansen et al., 1998; Cleeremans & McClelland, 1991; Elman, 1990). Alternative approaches include either (a) unfolding the recurrent net into a succession of forward projecting layers and thus treating each loop through the recurrent net as a distinct set of connections (back-propagation through time), or (b) keeping track in real time of the contribution of each connection during each iteration (real time recurrent learning) (Doya, 1995; Pearlmutter, 1995; Werbos, 1995). While theoretically feasible, these methods are not consistent with forward-running time and/or have memory and processing requirements that render them unwieldy for treating sequences with time-varying element durations in a straightforward and biologically feasible fashion (Werbos, 1995).

The current paper presents a temporal recurrent network model (TRN) that avoids this reduction of the temporal dimension. The technical difference comes from maintaining all connections into and within the recurrent network fixed, so that temporal variation in the role of a given recurrent connection over multiple time steps is no longer an issue for learning. The recurrent network thus encodes an internal state that is sensitive to the serial order of events, and the durations of their presentation and delays between them. Learning forms associations between states and appropriate responses via a simple associative learning rule (Dominey, 1995; Dominey et al., 1995). This approach is novel in that it is quite simple, yet it also provides a robust solution to the technical problem of encoding of serial as well as temporal structure in simulations of animal behaviour. Indeed the model was developed (Dominey et al., 1995) in order to explain electrophysiological recordings of neurons in the prefrontal cortex of the monkey in a temporal sequence learning task (Barone & Joseph, 1989).

The temporal recurrent network architecture is presented in Figure 1. Each of the components in Figure 1 corresponds to a 5×5 layer of leaky integrator "neurons", or units. In our simulations, each behavioural input

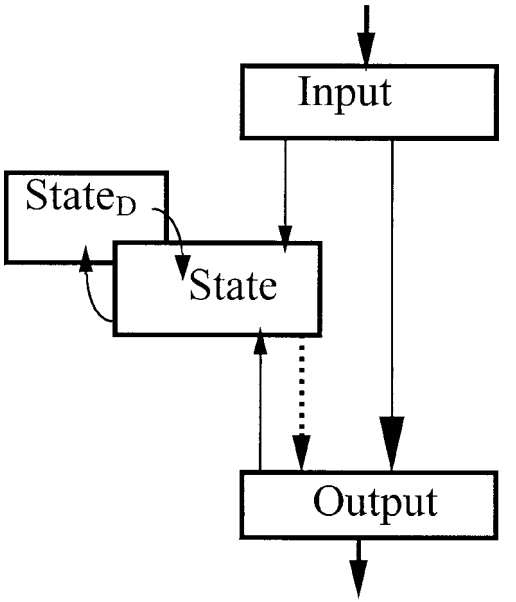

Figure 1. Temporal Recurrent Network (TRN) Sequence learning model based on recurrent state representation and associative learning. Each of the structures are 5 × 5 arrays of leaky-integrator units. Sequence elements are presented as activation of single units in the Input array. Responses are generated in Output. Output units are influenced by Input, and also by modifiable connections from State that form the associative memory. State is a recurrent network that encodes sequence state as a function of input from Input, response copy from Output, and recurrent self input from State$_D$. This network, State, generates a time varying sequence of internal states. These states become associated with Output activity for the successive responses by an associative learning mechanism that modifies State-Output connections. The model is implemented in neural Simulation Language (Weitzenfeld, 1991).

e.g., one of the 12 syllables in Saffran et al. (1996a) is mapped onto one of the 25 units in the Input layer. Likewise, a given behavioural response is generated as activation of one of the 25 Output units. The recurrent network, State, is an analogue to the primate frontal cortex which is characterised in part by its recurrent cortico-cortical connections (Goldman-Rakic, 1987). The associative memory linking State to Output finds its neurophysiological analog in dopamine modulated (Ljungberg, Apicella, & Schultz, 1991) NMDA receptor plasticity in corticostriatal synapses (Walsh & Dunia, 1993) that link frontal cortex to the major input nucleus of the basal ganglia that in turn provides motor and cognitive outputs to the cortex (Alexander, Delong, & Strick, 1986). The model has similarities with previous recurrent models (Elman, 1990; Pearlmutter, 1995; Pineda, 1989) with three important differences. First, as stated above, there is no

learning in the recurrent connections (i.e., those that project from $State_D$ to State and back), only between the State units and the Output units. Second, adaptation is based on a simple associative learning mechanism rather than back-propagation of error, or related error-gradient calculation methods (Pearlmutter, 1995). Third, in the temporal domain, (a) the computing elements are leaky integrators, and (b) simulation time steps are not tightly coupled to input, output and learning processing. That is, an input event can be specified to have a duration of any arbitrary number of time steps, and temporal delays between inputs can likewise be specified. Indeed, the experimenter's capability to specify the time delays between external events is an integral part of this model (Dominey et al., 1995).

Recurrent state representation

Equations (1.1) and (1.2) describe how the 5 × 5 unit layer State is influenced by external inputs from Input, recurrent inputs from $State_D$, and responses from Output. This recurrent state network was modelled after primate frontal cortex with its recurrent corticocortical connections (Goldman-Rakic, 1987), and allowed us to explain the electrophysiological encoding of visual space and sequential context (Dominey et al., 1995) recorded in neurons of the primate prefrontal cortex while monkeys performed learned movement sequences (Barone & Joseph, 1989). Equation (1.1) describes the leaky integrator, s(), corresponding to the membrane potential or internal activation of State. In Equation (1.2) the output activity level of State is generated as a sigmoid function, f(), of s(t). The term t is the time, Δt is the simulation time step, τ is the leaky integrator time constant. As τ increases with respect to Δt, the charge and discharge times for that leaky integrator increase. The unit of time in the simulations is referred to as a simulation time step or sts, and corresponds to a single update cycle of the simulation, and 5 milliseconds of simulated time.

$$s_i(t + \Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) s_i(t) + \frac{\Delta t}{\tau}\left(\sum_{j=1}^{n} w_{ij}^{IS}\, Input_j(t) + \sum_{j=1}^{n} w_{ij}^{SS}\, State_{Dj}(t) + \sum_{j=1}^{n} w_{ij}^{OS}\, Output_j(t)\right) \quad 1.1$$

1.2 $State(t) = f(s(t))$

The connections w^{IS}, w^{SS} and w^{OS} define the projections from units in Input, $State_D$, and Output to State. These connections are one-to-all, and are mixed excitatory and inhibitory, and do not change with learning. This mix of excitatory and inhibitory connections ensures that the State network does not become saturated by excitatory inputs, and also provides a source of diversity in coding the conjunctions and disjunctions of input, output and previous state information.

Recurrent input to State originates from the layer $State_D$. $State_D$ (Equations 2.1 and 2.2) receives input from State, and its 25 leaky integrator neurons have a distribution of time constants from 20 to 400 simulation time steps, while State units have time constants of 2 simulation time steps. This distribution of time constants in $State_D$ introduces a "damping" (hence the D) or low pass filtering that yields a range of temporal sensitivity similar to that provided by using a distribution of temporal delays (Kühn & van Hemmen, 1992).

2.1 $$sd_i(t + \Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) sd_i(t) + \frac{\Delta t}{\tau}(State_j(t))$$

2.2 $State_D = f(sd(t))$

An example of the temporal dynamics of this recurrent system is illustrated in Figure 2. A stimulus is provided as activation of one of the Input units (indicated in the lower trace) during 20 time steps, or 100 milliseconds, and the stimulus is then removed. In this sense, input events are not time-locked in a one-event-per-time-step fashion. We see that some State units are active during the activation of the Input unit, while others demonstrate activity that has a slower response, with activity changes taking place long after the stimulation is removed. The population of 25 State units thus provides a mechanism for coding the passage of time that can be used for learning to discriminate between different time intervals (Dominey, 1998b).

Associative memory

Thus, sequence context is continuously encoded in the recurrent network State. In order to use this context information to provide an appropriate response, the encoded context must become associated with the desired behavioural response in the output structure, Output. The required associative memory is implemented in a set of modifiable connections (w^{SO}) between State and Output, described in Equation (3). Each time a response is generated in Output, it is evaluated and the connections

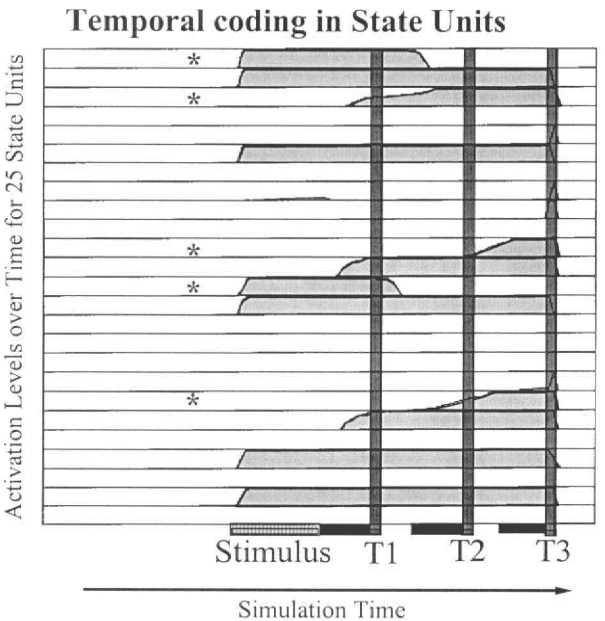

Figure 2. Illustration of the coding of temporal structure. In this simulation, a single unit in the Input layer was activated for 100 simulation time steps, illustrated on the temporal axis. With a short latency, some units in State became active. Others became active with a longer latency, even after the offset of the stimulus. The dark vertical lines indicate three temporal delays from the stimulus offset. The model was trained when presented with a go signal of three potential choices, to generate the correct responses in Output, depending on which one of the three delays occurred between the offset of the stimulus and the go signal. The units indicated with "*" differentiate between two of the three epochs, and together they encode the temporal delay structure such that the model can successfully learn the temporal discrimination task. (From Dominey, 1998b.)

between units encoding the current state in State, and the unit encoding the current response in Output are modified as a function of their rate of activation and learning rate R. R is positive for correct responses and negative for incorrect responses. Weights are normalised to preserve the total synaptic output weight of each State unit, thus avoiding saturation with extensive learning. Supervised and unsupervised learning with this rule are described later in this paper.

From a neurophysiological perspective, this associative memory is based on plasticity in the synapses linking cortical outputs to the striatum of the basal ganglia (Alexander et al., 1986). It has been demonstrated that when behavioural rewards (or events that predict rewards) occur, the neurotransmitter dopamine is released in the striatum (Ljungberg et al., 1992)

modulating the synpatic potentiation of NMDA receptors in corticostriatal synapses (Walsh & Dunia, 1993).

3. $w_{ij}^{SO}(t+1) = w_{ij}^{SO}(t) + R^{*}State_i {}^{*} Output_j$

The network output is thus directly influenced by the Input, and also by State, via learning in the w^{SO} synapses, as described in Equations (4.1) and (4.2). In Equation (4.2) the sigmoid output function f'() is the same as f() in Equations (1.2) and (2.2), and it additionally performs a winner-take-all function so that only one output neuron is active in the generated response.

4.1 $$o_i(t+\Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) o_i(t) + \frac{\Delta t}{\tau}\left(Input_i(t) + \sum_{j=1}^{n} w_{ij}^{SO}\, State_i(t)\right)$$

4.2 $Output = f'(o(t))$

The model has been used to explain the encoding of sequence context in neurons of the primate prefrontal cortex (Dominey & Boussaoud, 1997; Dominey et al., 1995) and learns complex sequences in reproduction (Dominey, 1995) and serial reaction time tasks (Dominey, 1988a, b). In addition, we have examined the ability of the model to learn and generalise in the domain of simple artificial languages in an effort to explore how the sequence processing capabilities of the primate cerebral cortex and basal ganglia might contribute to serial order processing in language (Dominey, 1997).

In general, parameters including the leaky integrator time constants and fixed connection strengths were tuned to maximise sequence learning performance and reproducibility of neurophysiological activity measured in the pre-frontal cortex. The parameters related to the recurrent State system were chosen in order to maximise the representation of sequential context over time, while also yielding a stable system. We have previously demonstrated that the model is fairly stable in the face of changes to the fixed parameters. Thus, the time constants in the $State_D$ units can vary by up to 100%, and the temporal delays between successive stimuli by up to 200% before producing significant impairments in sequence learning, depending on the sequence length and complexity (see Dominey, 1995).

Behavioural response modalities for the model

The combination of the context representation in State and the associative memory that binds these contexts to activity in Output allows

for at least two distinct types of sequence learning behaviour that we will use in the subsequent simulations. In the first form, unsupervised sequence learning is revealed as a reduction in the reaction times for elements in repeating sequences, with respect to reaction times for elements presented in a random order (Dominey, 1998a, b), as in the serial reaction time (SRT) paradigm of Nissen and Bullemer (1987). Reaction times (RTs) are measured as the delay between the onset of activation of a given Input unit, and the activation of the corresponding Output unit driven (in part) by the one-to-one connection from Input (Eqn. 4). This learning is unsupervised since the only possible response in Output is the one driven by the single activated Input unit. Recall that the response units in Output are leaky integrators whose response latencies are not instantaneous and depend on the strength of their driving sources (Eqn. 4.1). One source comes from the corresponding Input unit in a one-to-one mapping. This will activate the Output unit with some baseline RT. The other driving source for Output comes from State, which can change with learning in the w^{SO} Synapses (Eqn. 3). Since all responses in this unsupervised learning are correct, the learning rate R in Eqn. 3 is always positive. As learning occurs, RTs for elements in learned sequences will become reduced due to learning-specific influences of State on Output. This SRT learning in the model is understood in terms of three invariant observations: (1) During exposure to a repeating sequence, a given sub-sequence reliably generates the same pattern of neural activity in State. (2) This subsequence is reliably followed by a given element. (3) Learning results in strengthening of State-Output connections binding that pattern of activity in State to that sequence element in Output. These strengthened connections yield reduced reaction times for units in Output, for any element that is reliably preceded by the same sub-sequence, thus providing an SRT learning capability. This measure of learning will be used in the simulation of the Saffran et al. (1996a), Nazzi et al. (1998), and Marcus et al. (1999) results.

SRT learning thus relies on (a) regularities in the serial order of the sequence being reflected in the population of State units and (b) associations of these representations (patterns of activity in State) with activation of appropriate Output units via learning. The representation in State, however, does not rely on learning, as all connections into State including the recurrent connections, are fixed. Because of these connections, exposing the network to structurally different sequences yields different patterns or vectors of activity in the State units (see Dominey, 1998a, b). To the extent that sequences are related, their corresponding State vectors will be statistically related. The analysis of this relation will be used to demonstrate the State network's intrinsic and unsupervised

sensitivity to differences in the temporal structure of sentences originating from languages in different rhythm classes, similar to the sensitivity displayed by infants in the experiments of Nazzi et al. (1998).

This intrinsic capability to represent temporal or rhythmic structure can then be used to associate different output responses with different rhythmic classes. Presenting sentences from different rhythm classes to the network will yield different patterns of activity in State, dependent on the rhythm class. During training, the pattern or vector of activity in State for sentences for a given rhythm class will become linked, via associative learning (Eqn. 3), to a corresponding unit in Output. Likewise State vectors for sentences from a different rhythm class become linked to a different Output unit. In this supervised learning, correct responses result in the application of the associative learning rule (Eqn. 3) with a positive value of R to reinforce those connections contributing to the correct response, and incorrect responses use a negative value to weaken the connections contributing to the inappropriate response. Such learning yields a sequence discrimination or categorisation capability (Dominey, 1995), that we will apply to the same/different rhythm class discrimination based on the results of Nazzi et al. (1998). While such a discrimination protocol involves supervised learning, we note that the State vector analysis described above and the SET learning both demonstrate unsupervised sensitivity to prosodic structure.

A final form of sequence learning that we will not explore in this paper is explicit sequence reproduction, in which a sequence is presented to the model and the model must then reproduce that sequence, element by element. In this case, the model must for each element in the sequence choose from among all possible responses that are presented simultaneously in Input (the "go" signal), with the choice guided by learning. This involves the use of the associative memory to bind successive sequence contexts represented in State, to the successive sequence element responses produced in Output (Dominey, 1995; Dominey et al., 1995).

SENSITIVITY TO DISTRIBUTIONAL SERIAL STRUCTURE

In this section we describe the model's performance when exposed to the experiments of Saffran et al. (1996a) that demonstrated babies' sensitivity to the serial structure of sound sequences. In the simulation of Experiment 1, as in the original experiments, two sets of words were used for training in two different groups, A and B, and a third set of words were used for testing, as defined in Table 1. Recall that the training and testing are counterbalanced such that the first two words (Test A) in the test set are

TABLE 1
The Three-Syllable Words Presented in the Different Conditions of Experiment 1 and Experiment 2

	Group A	*Group B*	*Test*
Experiment 1	tu pi ro	da pi ku	Test A:
	go la bu	ti la do	tu pi ro, go la bu
	bi da ku	bu ro bi	Test B:
	pa do ti	pa go tu	da pi ku, ti la do
Experiment 2	pa bi ku	tu da ro	Test A:
	ti bu do	pi go la	pa bi ku, ti bu do
	go la tu	bi ku ti	Test B:
	da ro pi	bu do pa	tu da ro, pi go la

Note: that for both Experiments, Two Words (Test A) in the Test Condition are Words from Group A, and the Two Others (Test B) are Words from Group B. Adapted from Saffran et al. (1996b).

words from condition A and are non-words with respect to condition B. Likewise, the second two words in the test set (Test B) are words from condition B, and are non-words with respect to condition A. Thus, after exposure to condition A in an SRT learning context, we predict that in the test condition the model will respond with reduced reaction times for syllables in the Test A words with respect to RTs for syllables in the Test B words. The opposite should be the case after exposure to condition B.

Methods

For each of the two conditions A and B in Experiment 1, a pseudo-Random sequence of 180 words was produced from the original set of 4 words, thus yielding for each condition a sequence of 540 syllables. No particular word in the sequence immediately followed itself, and the transitional probabilities for syllables between words was .33 as in Saffran et al. (1996a). Each of the 12 syllables was mapped onto one of the 25 Input units of the model, leaving the remaining 13 Input units unused. In order to study the stable population behaviour of the network, we report results from a population of 10 model "subjects" created by using different random number generator seed values in initialising the connections w^{SO}, w^{IS}, w^{SS} and w^{OS}. Five subjects were exposed to one repetition of the 540 element sequence A, and the other five to the 540 element sequence B. Both groups were then exposed to the same test sequence material, and RTs were obtained separately for occurrences of elements in the Test A and Test B words in the test set.

As noted by Saffran et al. (1996a) this problem could be solved by learning the serial order of events presented during the two minute training, and then recognising these sequences during testing, since the

non-words had never been heard during training. In order to determine if babies were also sensitive to the relative frequencies of presentations of different sequences, in Experiment 2, rather than using non-words that had never been heard, Saffran et al. (1996a) used part-words, constructed from syllables that occurred at word boundaries in the training set. Thus, the babies had heard these part-words, but at a reduced frequency with respect to the words. The simulation thus took this into account, using the stimuli of experiment 2 of Saffran et al. (1996a) as shown in Table 1, and otherwise following the method described for Experiment 1, including the restriction that no word followed itself and the transitional probabilities for syllables between words was .33.

Results and discussion

In the simulation of Experiment 1, the response times in Group A were reduced for the Test A words of the test set, and increased for the Test B words (Figure 3A). The RT can be considered a measure of novelty, thus the Test A words of the test set were processed as being more familiar, while the Test B words processed as being novel. In contrast, for Group B, as predicted, the opposite was seen. These observations were confirmed by a 2 × 2 repeated measures ANOVA in which the between subjects variable was Group (A, B), the within subjects variable was Test words (Test A, Test B) and the dependent variable was the response time for the syllables in the test words. The effect for Group was not significant [$F(1, 8) = 0.18$, $p = .68$] with no overall difference between the two groups. Likewise, the effect for Test words was also not significant [$F(1, 8) = 0.08$,

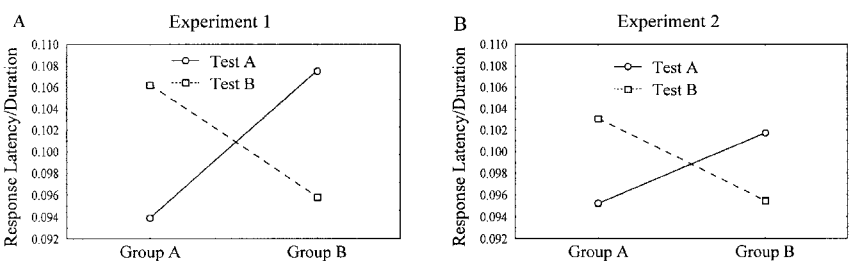

Figure 3. (A) Simulation of Experiment 1 of Saffran et al. (1996b). Groups A and B were trained on different phoneme strings (see Table 1). Test A words were heard by Group A in training, and not by group B, with the opposite for the Test B words. Both groups show reduced response duration for test stimuli that they had been exposed to during training. (B) Simulation of Experiment 2 of Saffran et al. (1996b). Test A words of test stimuli were high frequency words for Group A, and low frequency words for Group B, with the opposite for the Test B words. Again, both groups show reduced response duration for test elements that they were more exposed to during training.

$p = .78$] with no overall difference in the responses to the Test A or Test B words of the test set. However, the Interaction between these two factors was significant [$F(1, 8) = 125.2$, $p < .0001$] as the group A subjects responded preferentially to the Test A words, and group B subjects to the Test B words.

Similarly, for the simulation of Experiment 2 seen in Figure 3B, Group A produces reduced response times for the Test A words of the test set, and increased response times for the Test B part-words. That is, the Test A words of the test set were "recognised" as being more familiar, while this was not the case for the Test B part-words. In contrast, for Group B, as predicted, the opposite was seen. These observations were confirmed by a 2×2 repeated measures ANOVA in which the between subjects variable was Group (A, B), the within subjects variables were Test words (Test A, Test B), and the dependent variable was the response time. The effect for Group was not significant [$F(1, 8) = 0.04$, $p = .86$] with no overall difference between the two groups. Likewise, the effect for Test words was also not significant [$F(1, 8) = 0.17$, $p = .69$] with no overall difference in the responses to the first two or last two words of the test sequence. However, again the Interaction between these two factors was significant [$F(1, 8) = 12.8$, $p < .01$] as the group A subjects responded preferentially to the Test A words, and group B subjects to the Test B words. These observations demonstrate that like babies, the model is sensitive to the serial order of events, and to their distributional frequency of occurrence.

SENSITIVITY TO TEMPORAL STRUCTURE

The results of the Saffran et al. (1996a) experiment simulations verify that the model is sensitive, as are babies, to statistical regularities in the serial structure of sound sequences, one of the major phonological information sources that appear to be treated by babies. In this section we describe the simulation studies of the experiments from Nazzi et al. (1998) that demonstrate sensitivity to temporal structure.

The study described here is based on the hypothesis that infants represent the speech stream as a sequence of vowels separated by possibly unanalysed intervals, i.e. consonants. This is because "vowels carry most of the energy in the speech signal, they last longer than most consonants, and they have greater stability. They also carry accent and signal whether a syllable is strong or weak" (Mehler et al., 1996, p.112). Furthermore, newborns seem to pay more attention to vowels than to consonants (Bertoncini et al., 1988), and to be sensitive to the number of syllables (hence vowels) in a string, regardless of syllable structure or weight (Bertoncini et al., 1995; Bijeljac-Babic, Bertoncini, & Mehler, 1993; van Ooyen et al., 1997). We are not claiming that infants can perfectly segment

speech into precise consonant and vowel durations, but rather, that speech rhythm perception in the infant is primarily based on the extraction of temporal regularities over consonantal and vocalic intervals.

Taking this hypothesis seriously, Ramus et al. (in press) have shown that a simple segmentation of speech into vocalic and consonantal intervals accounts for the different types of linguistic rhythm, and have therefore proposed a model of language discrimination in newborns that rests on such a consonant/vowel segmentation of speech. This is consistent with another study showing that the CV alternation is all that is needed by adult subjects to discriminate between different types of rhythm (Ramus & Mehler, 1999), and with psychophysical and neurophysiological evidence that rhythm perception and production at this time-scale may rely on interval-based rather than beat-based processes (Ivry & Hazeltine, 1995; Hooper, 1998). Given this converging evidence, it seems reasonable to assume here that speech rhythm perception is primarily based on the extraction of temporal regularities over consonantal and vocalic intervals.

Thus the present simulations took as input speech that was pre-segmented into consonant (C) and vowel (V) durations, sampled and coded at 5 ms intervals, e.g., V – 0.110, C – 0.060, V – 0.065, C – 0.075, ... Two Input units were used, one for C and one for V. These units were activated in C–V–C–V . . . sequences that respected the coded intervals. The main idea is that if indeed the model is sensitive to temporal structure (Dominey, 1998a, b) then presentation of these . . . C–V–C–V–C–V . . . sequences that have identical serial structure but different temporal structures should result in different vectors of activity in the 25 State neurons. Thus, sequences derived from sentences from the same rhythm class should yield similar State vectors, while those from different rhythm classes should be systematically different.

Method

In order to explore the inherent sensitivity of the recurrent State system to temporal variations in C–V sequences without training (supervised nor unsupervised) we exposed the network to 10 3-second sentences from each of five languages (coded as C–V-C–V . . . sequences as described above), and recorded the State vector resulting from each of these sentences. We then tested whether these State vectors could be correlated with the languages from which they originated. Table 2 (State-Class Correlation) indicates that for the language pairs in which the required discrimination was between languages from different rhythm classes, there was a significant correlation between State vector activity and language, and the opposite for discriminations between languages from the same rhythm class. Given this indication that the recurrent State network is sensitive to

TABLE 2
Human and Simulated Performance on Language Discrimination Experiments 1–3

	Languages	*Rhythm Class*	*Babies*	*State-Class Correlation*	*Performance*
Exp. 1	Eng vs. J	Different	$p < .01$	$R^2 = .75$, $p < .001$	78%, $p < .001$
Exp. 2	Eng vs. Dutch	Same	$p = .16$	$R^2 = .17$, $p = .87$	52%, $p = .86$
Exp. 3a	E+ D vs. S+ I	Different	$p < .01$	$R^2 = .49$, $p < .001$	73%, $p = .001$
Exp. 3b	E+ S vs. D+ I	Same	$p = .20$	$R^2 = .17$, $p = .34$	55%, $p = .42$

Languages: English (Eng), Dutch (D) (Stress-Timed), Spanish (S), Italian (I) (Syllable-Timed), Japanese (J) (Mora-Timed). Rhythm class: According to Pike (1945) and Abercrombie (1967). Babies: Discrimination Performance in Infants from Nazzi et al. (1998). State-Class Correlation: Measure of Degree to which Model's State Vector Contents for a Given Sentence can Predict the Sentence's Rhythm Class. Performance: Percent Correct in Rhythm Class Discrimination for the Model.

rhythm class differences, we then exposed the model to the three experiments of Nazzi et al. (1998).

In order to study the stable population behaviour of the network, we report results from a population of 10 model "subjects" created by using different Random number generator seed values in initialising connections w^{SO}, w^{IS}, w^{SS} and w^{OS}. Simulations assessed in three conditions the ability to discriminate between sentences from two languages, in terms of the percentage of correct classifications, where 50% represents chance performance. The *Naive* condition tested discrimination on 10 3-second sentences (20 in Exp. 3) from each of the two languages with no learning. The *Training* condition tested discrimination on 10 new sentences (20 in Exp. 3) from each language, using the associative learning rule. The *Generalisation* condition then tested the trained model (with learning now inactivated) on the original sentences from the Naive condition, to assess the generalisation of learning acquired during the Training condition (Figure 4). Four speakers per language (two in training, 2 in generalisation) were used to ensure that learning did not depend on a particular speaker's characteristics. Simulation performance in the *Generalisation* condition is compared with that of infants from Nazzi et al. (1998) in Table 2 ("Performance" and "Babies", respectively).

Finally, as this generalisation protocol requires a supervised learning during the training phase, we wanted to verify that the model could perform a non-supervised learning as in the protocol employed by Nazzi et al. (1998). We thus simulated Experiment 1, based on the habituation protocol using 10 model subjects. During the habituation phase, the model

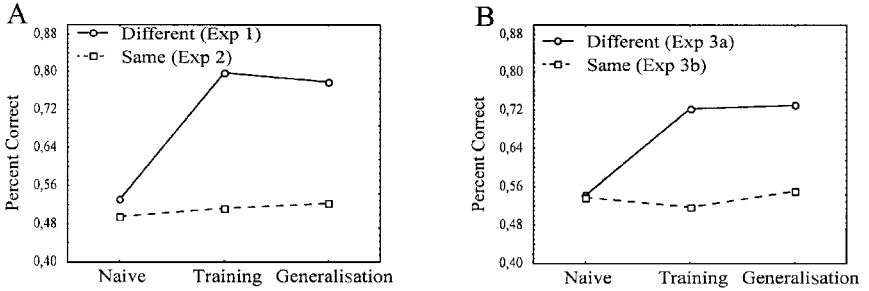

Figure 4. (A) Simulation based on Experiments 1 and 2 of Nazzi et al. (1998), for discrimination between languages from Different rhythm classes (English and Japanese), and languages from the Same rhythm class (English and Dutch). The model is sensitive to the prosodic differences only for the Different discrimination case for training and generalisation to new sentences. (B) Simulation of Experiments 3a and 3b of Nazzi et al. (1998), for mixtures of languages yielding discrimination between two Different rhythm classes (English + Dutch vs. Italian + Spanish), and for mixtures of languages yielding discrimination between the Same mixed rhythm class (English + Spanish vs. Dutch + Italian). The model is sensitive to the prosodic differences only for the Different discrimination case for training and generalisation to new sentences, indicating that the sensitivity is not at the level of the given language, but at the level of the rhythm class.

was exposed to 10 CV-resynthesised sentences, twice each, in the target language. After the presentation of each sentence, a response to a single probe stimuli (the same for all sentences) was elicited. Unsupervised learning should construct the association between prosodic structure of the target sentences represented in State, and the activated probe response unit. After this habituation, the model was then presented with new CV coded sentences from different speakers in the target language, and in a different language. If learning occurred during the habituation, reaction times to the sentences in the same language should be reduced with respect to those in the different language.

Results and discussion

Experiment 1 tested the ability to discriminate between English and Japanese which come from different rhythm classes. Experiment 2 tested the ability to discriminate between English and Dutch, which are both in the same rhythm class. As seen in Figure 4A, naive model performance is near chance level for both cases. For discriminations that contrast sentences from one rhythm class with sentences from a different rhythm class (Exp. 1), the model learns, and can then generalise this learning to new sentences uttered by new speakers. Learning and generalisation fail for discriminations that contrast sentences from one rhythm class with

sentences from the same rhythm class (Exp. 2). That is, between-class contrasts succeed, while within-class contrasts fail.

These observations were confirmed by a 2 × 2 repeated measures ANOVA in which the between subjects variable was Rhythm Class (Exp. 1—Different/Exp. 2—Same), the within subjects variable was Condition (Naive, Training, Generalisation) and the dependent variable was the Performance in terms of percent correct in discrimination. There was a significant effect for Rhythm Class [$F(1, 18) = 45.5$, $p < .0001$] with improved performance for Different vs. Same comparisons. The effect for Condition was also significant [$F(2, 36) = 20.1$, $p < .0001$] with improved performance in Training and Generalisation with respect to Naive conditions. Finally, the interaction was significant [$F(2, 36) = 14.3$, $p < .0001$] as these Condition effects were dependent on the Rhythm Class.

While Experiments 1 and 2 examined discrimination between single language pairs, i.e., English–Japanese and English–Dutch, the hypothesis that children are discriminating not languages but rhythm classes allows for a stronger prediction. That is, children should be capable of discriminating a mixture of sentences from languages in one rhythm class (e.g., English and Dutch) from a mixture of languages in a different rhythm class (e.g., Spanish and Italian). This is the basis of Nazzi et al.'s (1998) Experiment 3. Thus, in Exp. 3a, babies were habituated with sentences from English and Dutch and then tested either with new English and Dutch sentences, or with sentences from Spanish and Italian, with appropriate controls in the opposite sense. In Exp. 3b, babies were habituated with sentences from two different rhythm classes (e.g., English and Spanish) and tested with sentences from these same rhythm classes (e.g., Dutch and Italian). In Exp. 3a babies could discriminate between the two groups of languages, while in Exp. 3b they failed. This argued that under these conditions, babies are not sensitive to specific language differences per se, but rather to the rhythm classes to which these languages belong.

In simulation parts A and B the model was thus exposed to 10 sentences from each of the two target languages (20 sentences in all), and then tested on 10 sentences from each of the two test languages. As seen in Figure 4B, the model, like the babies, is capable of discriminating between English mixed with Dutch, vs. Spanish mixed with Italian. Also, like the babies, it fails to discriminate English mixed with Spanish vs. Dutch mixed with Italian.

These observations were confirmed by a 2 × 3 repeated measures ANOVA in which the between subjects variable was Rhythm Class (Exp. 3a—Different/ Exp. 3b—Same), the within subjects variable was Condition (Naive, Training, Generalisation) and the dependent variable was the Performance in terms of percent correct in discrimination. There was a significant effect for Rhythm Class [$F(1, 18) = 35.0$, $p < .0001$] with

improved performance for Different vs. Same comparisons. The effect for Condition was also significant [$F(2, 36) = 19.2$, $p < .0001$] with improved performance in Training and Generalisation with respect to Naive. The Interaction was significant [$F(2, 36) = 15.1$, $p < 0.001$] as these Condition effects were dependent on the Rhythm Class. These results indicate that the model can learn, with supervision, to discriminate between languages from different rhythm classes.

In the habituation simulations, we wanted to verify that the model could discriminate between different rhythm classes in unsupervised conditions in the protocol of Nazzi et al. (1998). Following the design of Nazzi et al.'s Experiment 1, the model was exposed (or habituated) to target sentences (English or Japanese) and then tested with new sentences, in both the same language, and in a different language. Each of the 10 model subjects was separately tested with English and Japanese as the habituation/target language. We analysed the models' reaction times for the new sentences with the assumption that new sentences in the target language will benefit from the habituation exposure and thus demonstrate reduced reaction times. As seen in Figure 5, RTs are indeed smaller for new sentences from the same language as those used in the habituation training, with increased RTs for sentences from a different language. This observation was confirmed in a repeated measures ANOVA in which sentence type (Same vs. Different) was the within subjects variable, and RT was the dependent

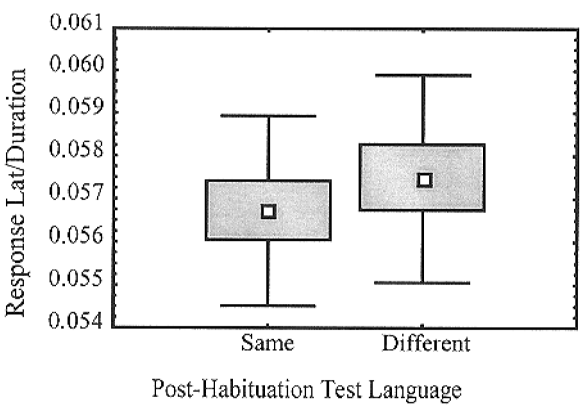

Figure 5. Simulation based on Experiment 1 of Nazzi et al. (1998), following the unsupervised learning protocol. Subjects are habituated to a target language, then tested with sentences from the same, or from a different language. The model is sensitive to the prosodic differences between English and Japanese, demonstrating significantly increased response durations only for sentences in the post-habituation testing that were in a different language from those used in the habituation.

variable. The effect of sentence type was reliable [$F(1, 9) = 11.06$; $p < 0.01$], with RTs reduced for Same vs. Different sentence types.

Based on these simulation results, and those summarised in Table 2, we can conclude that like the babies, the model is sensitive to the rhythmic structure of sentences. This sensitivity is related to the rhythmic structure at the class level such that languages in different rhythm classes can be distinguished, while languages within the same rhythm class cannot be distinguished. The model can simulate baby-like sensitivity to temporal structure, including unsupervised learning as in the Nazzi et al. (1998) experiments. These simulations leave open the question of how this sensitivity translates to the observed behaviour, though we note that sucking rate in the infant, and response time in the model both increase for novel stimuli.

SENSITIVITY TO ABSTRACT STRUCTURE AND THE ABSTRACT RECURRENT NETWORK (ARN)

In this section we explore the sensitivity to abstract structure as displayed by infants in the study of Marcus et al. (1999). We recall that Marcus stresses the point that standard sequence learning models will fail in this task, since they represent sequences in terms of their constituent elements, and thus cannot transfer knowledge to new sequences made up of entirely unfamiliar constituents. We have previously explored this type of transfer between isomorphic sequences such as ABCBAC and DEFEDF that have different surface structures but a common abstract structure 123213 (Dominey, 1997; Dominey et al., 1998). We demonstrated that the TRN fails to learn and transfer this abstract structure to new isomorphic sequences, and that in order to do so, the model must first be augmented with a working or short term memory (STM) of the 5 (7 ± 2) preceding elements, described in Eqn. 5, and illustrated schematically in Figure 6.

After each response in Output to an element presented in Input, the five STM structures (each a 5×5 array) are updated to reflect the n–1st to n–5th previous responses, respectively.

5.1 for i = 5 to 2, STM(i) = STM(i– 1)

5.2 STM(1) = Output

A recognition function then compares current response in Out with the STM contents, in Eqn. 6, to detect whether the current element is a repetition of a previous one, as indicated in Figure 5. In Eqn. 6 the "*" performs a pointwise multiplication yielding a non-zero result only if two corresponding elements in Output and one of the STM arrays are active,

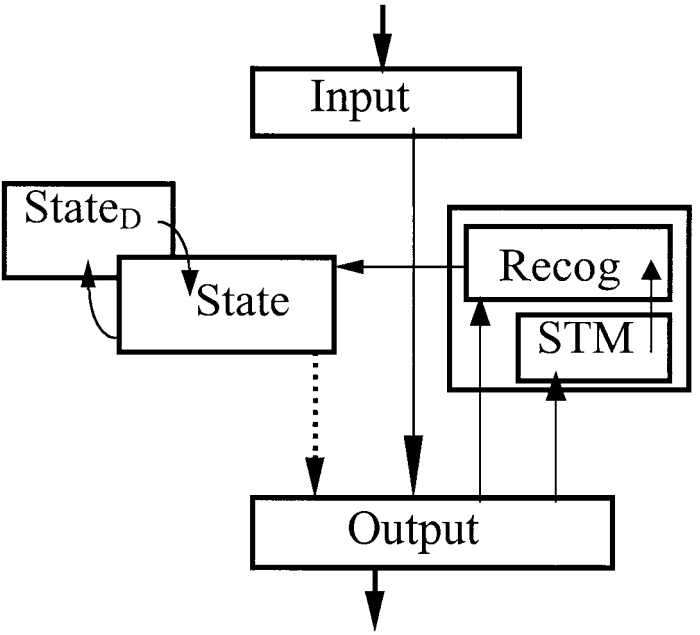

Figure 6. Updated Abstract Recurrent Network (ARN) model to include capabilities for abstract structure processing. "STM" is a short-term or working memory of the five previous stimuli (e.g., words). "Recog" is a 6 unit recognition function that compares the current stimulus with the contents of STM. If the current stimulus matches STM element (n), the unit Recog(n) is activated. If no match is made, then Recog(6) is activated, indicating a unique or non-repeating element. This recognition-related activity from Recog is provided as input to State. Thus State represents the "abstract" rather than the "surface" structure of the sequences, such that "je de de" and "wi di di" now have the same representation, ABB.

i.e., only if the current sequence element is the same as one of the previous five elements in the sequence.

6. $\text{Recognition}_i = \text{STM(i)} * \text{Out}$

In this manner sequences can be coded in the modified abstract recurrent network (ARN) in terms of their internal repetitive structure. Thus ABCBAC and DEFEDF are both encoded U U U n–2 n–4 n–3 where U indicates a unique or non-repeating element, and n–2 indicates a repeat of the element 2 places back, and so on. This abstract representation in Recognition then feeds the State mechanism, and replaces the input from the Input units as indicated in Eqn. 1.1*, thus allowing a representation of sequential context at this abstract level.

$$s_i(t + \Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) s_I(t) + \frac{\Delta t}{\tau}\left(\sum_{j=1}^{n} w_{ij}^{SS}\, StateD_j(t) + \sum_{j=1}^{n} w_{ij}^{RS}\, Recog_j(t)\right) \quad 1.1^*$$

In this context, the pattern of activity in State resulting from the presentation of ABA-grammar sequences DCD and HAH should thus be identical.[1] This identical representation in State can be demonstrated by running the following simulation. During a habituation period, after each sentence is presented, a single test stimulus is presented in Input (the same one for all sentences) and the RT to this stimulus in the one and only corresponding unit in Output is measured. As a result of learning in the State-Output connections the patterns of activity in State units that are generated by exposure to ABA-grammar sentences become associated with activation of the corresponding Output unit via the associative memory, thus yielding reduced RTs. Testing with new ABA sentences will generate similar State vectors and will thus exploit this learning and activate the Output unit with a reduced RT. Sentences not consistent with ABA will generate different State vectors with resulting increased RTs for the Output activation.

Thus, the representation capability in State is fixed, and what is learned is the associations between the ABA-driven patterns of activity in the State units, and the response in the Output unit. Reduced RTs in testing with new ABA-grammar sentences thus reflects similarity to the training material. In these simulations, we will first consider the behaviour of the described ARN, and then the TRN.

In the simulation of Experiments 1 and 2, as in the original experiments, two sets of sentences from two grammars ABA and ABB were used for training in two different groups. A third set was used for testing the transfer of the acquired knowledge to new sentences, as defined in Table 3. Recall that the training and testing are counterbalanced such that two sentences in the test set are consistent with the grammar ABA while the other two sentences are consistent with grammar ABB. Thus, after exposure to condition ABA, we predict that in the test condition learning should be reflected as reduced reaction times for the ABA-grammar sentences of the test condition with respect to RTs for the ABB-grammar sentences. The opposite should be the case after exposure to condition ABB.

[1] In Dominey et al., 1998, this architecture was further modified so that State could influence the modulation of STM contents into the Output, so that the model could predict successor elements in new isomorphic sequences that followed a learned abstract structure, and we refer the interested reader there for more details.

Methods

In Marcus et al. (1999), Experiment 2 controlled for a possible confound in Experiment 1 due to voicing. Here such cues are irrelevant, thus we directly simulated Experiment 2, and we will refer to it as Experiment 1-2. For each of the two training conditions ABA and ABB in Experiment 1-2, a pseudo-random sequence of sentences was produced from three repetitions of the original set of 16 sentences that each consisted of three words (see Table 3), thus yielding for each condition a sequence of 144 words (16 sentences $\times$ 3 words/sentence $\times$ 3 repetitions). Each of the 12 words was mapped onto a different Input unit of the model, leaving 13 Input units unused. In Marcus et al. (1999), each sentence was separated from the next by a 1.2–1.5-second pause that effectively "reset" the

TABLE 3
The Three-Word Sentences Presented in the Different Conditions of Experiment 1-2 and Experiment 3

	Training Group A:	*Training Group B:*	*Test:*
Experiment 1-2	*Training ABA*	*Training ABB*	*Test ABA vs. ABB*
	le di le, le je le, le li le, le we le,	le di di, le je je, le li li, le we we,	ABA: ba po ba ko ga ko
	wi di wi, wi je wi, wi li wi, wi we wi,	wi di di, wi je je, wi li li, wi we we,	ABB:
	ji di ji, ji je ji, ji li ji, ji we ji,	ji di di, ji je je, ji li li, ji we we,	ba po po ko ga ga
	de di de, de je de, de li de, de we de	de di di, de je je, de li li, de we we	
Experiment 3	*Training AAB*	*Training ABB*	*Test AAB vs. ABB*
	le le di, le le je, le le li, le le we,	le di di, le je je, le li li, le we we,	ABA: ba ba po ko ko ga
	wi wi di, wi wi je, wi wi li, wi wi we,	wi di di, wi je je, wi li li, wi we we,	ABB:
	ji ji di, ji ji je, ji ji li, ji ji we,	ji di di, ji je je, ji li li, ji we we,	ba po po ko ga ga
	de de di de de je, de de li, de de we	de di di, de je je, de li li, de we we	

Note: For both Experiments, Two Words in the Test Condition Sentences Follow the Abstract Structure from Group A, and Two Follow the Abstract Structure from Group B. Each Word Corresponds to One Unit in the 25 Unit Input Layer. Adapted from Marcus et al. (1999).

processing so that the previous sentence did not merge into, or interfere with the current one. We thus begin the processing of each new sentence with the State units and the contents of the working memory cleared of residual activity from the previous sentence to simulate this effect. In the transfer testing with new sentences made of words not presented during training, the model was exposed to two repetitions each of the two test sentences of type ABA and the two sentences of type ABB (see Table 3). We report results from 2 populations of 10 model "subjects" created by using different Random number generator seed values in initialising the connections w^{SO}, w^{IS}, w^{SS} and w^{OS}. One group of 10 subjects was exposed to one repetition of the 144 element sequence ABA, and the other to the 144 element sequence ABB. Both groups were then exposed to the same test sentence material, and RTs were obtained separately for occurrences of sentences of type ABA and ABB in the test material.

As noted by Marcus et al. (1999) discrimination between ABA and ABB could be achieved by learning to detect reduplication or doubling present in ABB but not ABA. In order to verify that babies were not simply exploiting this difference, in Experiment 3 the two grammars were AAB and ABB, both of which have reduplication. The simulation of Experiment 3 thus took this into account, using the stimuli of Marcus et al. (1999) as shown in Table 3, and otherwise following the method described for Experiment 1-2.

Results and discussion

In the simulation of Experiment 1-2, using the ARN that has been demonstrated to be sensitive to abstract structure, the response times in Group ABA were reduced for test sentences of type ABA, and increased for ABB sentences (Figure 7A). As an RT increase can be considered a measure of novelty, ABA sentences made of new words were processed as being more familiar, while ABB sentences were processed as being novel. In contrast, for Group ABB, as predicted, the opposite was seen. Likewise, in the simulation of Experiment 3 (Figure 7B), the AAB group showed reduced RTs for new AAB vs. ABB sentences, and the ABB group demonstrated the opposite effect. Thus it appears that for both Experiments 1-2, and Experiment 3 the simulations replicate the babies' performance in learning the abstract structure. Recall that test sentences were made of words not used in training, so any performance transfer from training to testing had to reflect knowledge of the abstract rule or grammar.

These observations for Experiment 1-2 were confirmed by a repeated measures ANOVA in which the independent variables were Training condition (ABA, ABB), Transfer condition (ABA, ABB), and Transfer

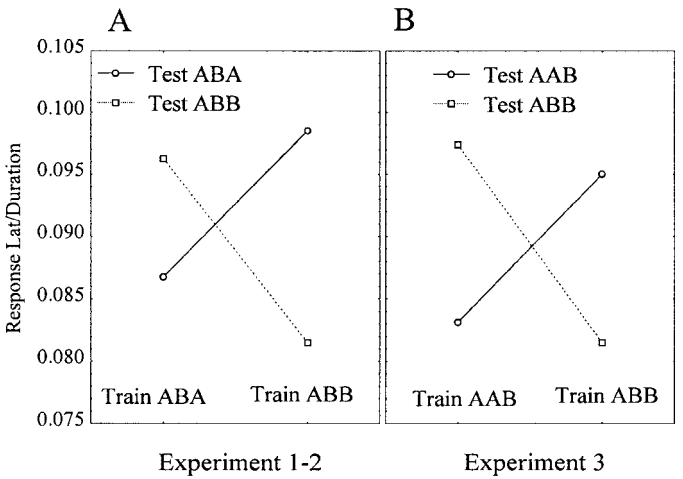

Figure 7. Simulation of Experiments 1-2, and 3 from Marcus et al. (1999) using the updated Abstract Recurrent Network (ARN) of Figure 5. (A) Simulation of Experiment 1-2. Two model Groups were trained on sentences of type ABA and ABB respectively (see Table 3). Two test stimuli were sentences of type ABA and two were of type ABB. Both groups show reduced response duration for test sentences of the type that they had been exposed to during training. (B) Two model Groups were trained on sentences of type AAB and ABB respectively (see Table 3). Two test stimuli were sentences of type AAB and two were of type ABB. Both groups show reduced response duration for test sentences of the type that they had been exposed to during training, as observed in the human infant results of Marcus et al. (1999) for abstract structure processing.

sentence (1–4) and the dependent variable was the response time. The effect for Training was not significant [$F(1, 9) = 0.23$, $p = .63$] with no overall difference between the two groups. Likewise, the effect for Transfer was also not significant [$F(1, 9) = 2.26$, $p = .16$], nor was the effect of the four specific test sentences within the two types significant [$F(3, 27) = 0.42$, $p = .74$]. However, the Interaction between Training and Transfer was significant [$F(1, 9) = 46.6$, $p < .0001$] as the group ABA subjects responded preferentially to the ABA transfer sentences, and group ABB subjects responded preferentially to the ABB transfer sentence.

Similarly, for the simulation of Experiment 3 seen in Figure 7B, Group AAB produces reduced response times for the AAB transfer sentences, and increased response times for the ABB transfer sentences. That is, AAB transfer sequences were "recognised" as being more familiar, while this was not the case for the ABB sentences. In contrast, for Group ABB, as predicted, the opposite was seen. These observations for Experiment 3 were confirmed by a repeated measures ANOVA in which the independent variables were Training condition (AAB, ABB),

Transfer condition (AAB, ABB), and Transfer sentence (1–4) and the dependent variable was the response time. The effect for Training was not significant [$F(1, 9) = 0.30$, $p = .59$] with no overall difference between the two groups. Likewise, the effect for Transfer was also not significant [$F(1, 9) = 0.02$, $p = .89$], nor was the effect of the four specific test sentences within the two types significant [$F(3, 27) = 0.54$, $p = .65$]. However, the Interaction between Training and Transfer was significant [$F(1, 9) = 10.32$, $p = .011$] as the group AAB subjects responded preferentially to the AAB transfer sentences, and group ABB subject to the ABB transfer sentence.

These results indicate that the ARN model simulates the babies' sensitivity to abstract structure representations in which sentences such as "le di di" and "we je je" are represented as being related to each other as ABB, and unrelated to sentences such as "ba ba po". In Figure 8 we display the performance of the TRN on these same tasks. While the ARN

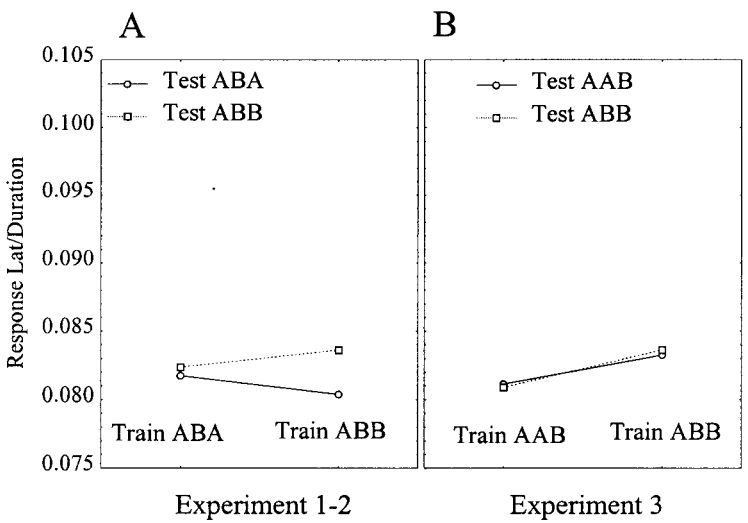

Figure 8. Simulation of Experiments 1-2, and 3 from Marcus et al. (1999) using the Temporal Recurrent Network (TRN). (A) Simulation of Experiment 1-2. Two model Groups were trained on sentences of type ABA and ABB respectively (see Table 3). Two test stimuli were sentences of type ABA and two were of type ABB. Both groups fail to show reduced response duration for test sentences of the type that they had been exposed to during training. (B) Two model Groups were trained on sentences of type AAB and ABB respectively (see Table 3). Two test stimuli were sentences of type AAB and two were of type ABB. Both groups fail to show reduced response duration for test sentences of the type that they had been exposed to during training. The TRN model thus fails to simulate the human infant results for abstract structure processing of Marcus et al. (1999).

model displayed clear performance differences on the transfer material depending on the training condition, this does not appear to be the case for the TRN. These observations were confirmed for Experiments 1-2, and Experiment 3 by the same ANOVAs as in the ARN analysis above. None of the main effects were significant, nor were any of the interactions. Most importantly, the critical Training $\times$ Testing interactions that would indicate learning of the abstract structure for Experiment 1-2 [$F(1, 9) = 0.17$, $p = .68$] and for Experiment 3 [$F(1, 9) = 1.6$, $p = .23$] were not significant.

The previous two sections demonstrated that the TRN adequately represents regularities in serial order, and in the temporal or rhythmic structure. Here we see that this same model fails to represent abstract structure. This abstract structure can be represented, however, by a modified version of this model that represents relations between repeating elements, rather than the elements themselves (Dominey, 1997; Dominey et al., 1998). These results are in agreement with the conclusions of Marcus et al. (1999) that the infants possess at least two distinct learning mechanisms that can contribute to language acquisition. A mechanism such as our TRN represents statistical regularities, and a dissociable mechanism such as our ARN represents, extracts and generalises abstract rules.

GENERAL DISCUSSION

Babies are remarkably sensitive to regularities in the serial order of acoustic events in language (e.g. Saffran et al., 1996b), their temporal structure (e.g., Nazzi et al., 1998), and their abstract structure (e.g., Marcus et al., 1999). While these observations are supported by empirical data, they leave open the question of the computational and underlying neurophysiological processes or mechanisms responsible for this sensitivity.

We previously demonstrated that our temporal recurrent network (TRN) based on the primate frontostriatal system (Dominey et al., 1995) is a plausible model of the neural circuitry underlying the capacity of monkeys to learn, generalise and discriminate between specific serial structures in which the temporal structure is fixed (Dominey, 1995, 1997). Here we showed that this very model can exhibit the same abilities when confronted with a different, but formally similar sort of input, that is, a string of syllables constituting a miniature artificial language (Saffran et al., 1996a). Since the sequences of syllables we used have the same formal structure as the visuomotor sequences for which the model was designed, it is hardly surprising that the TRN exhibits the same abilities

for both types of input. This merely helps us to demonstrate that it is also a plausible model of how the ability of infants to extract distributional regularities could be implemented. Of course, other models are available. A Simple Recurrent Network (Cleeremans & McClelland, 1991; Elman, 1990) can certainly do the job as successfully, as suggested by the work of Christiansen et al. (1998).[2] Beyond bold generalisations from animal to human models, no neurophysiological evidence is currently available to help determine which model is the most plausible implementation of infants' abilities.

However, the TRN was also shown to be sensitive to certain temporal regularities in sequences that share the same serial structure (Dominey, 1998a, b). The current study extends this result to the particular temporal regularities underlying the global rhythmic properties of natural languages (Nazzi et al., 1998; Ramus et al., in press). The ability of the TRN to represent and learn temporal regularities is interesting in the light of the difficulty (Werbos, 1995) of simple recurrent networks to do so. Their use of learning in the recurrent connections typically enforces a requirement that at each time step a new input is processed and the output evaluated, with no natural way to process relative temporal durations of sequential events and the delays between them without excessive computational and/or memory requirements (Werbos, 1995). This limitation can effectively be overcome as, for example, when Christiansen et al. (1998) used an SRN to detect regularities signalling word boundaries, they used special symbols to represent prosodic regularities in the input. For instance, they used a special input symbol in order to represent pauses at utterance boundaries, and another one to signal whether a phoneme belonged to a stressed syllable or not. This could be considered an intermediate coding level where acoustic events are discretised into a few symbols. Of particular relevance with respect to the current study, however is their use of symbolic coding for temporal/prosodic structure. The TRN is more parsimonious in this respect, since in our simulation of Nazzi et al. (1998), temporal structure is simply represented by actual delays during and between sequence elements, and stressed syllables are signalled by

[2] See also Perruchet and Vinter (1998), who have recently argued that such performance can result from a relatively simple associative learning mechanism, in which the internal representations for repeated percepts become strengthened through their repetition, while representations for infrequent percepts are weakened. Their model thus maintains a list of candidate percepts and continuously compares the next encountered percept with this list, adding novel percepts to the list, strengthening matching percepts and weakening the others. After exposure to a training corpus, the model reliably extracts the regular repeating words as the strongest surviving percepts. One defining characteristic of this model is that the management of the percept list relies on an algorithm-like assembly of specialised rules.

their lengthening.[3] In other words the temporal structure, which is the object of interest here, is realistically represented as durations of events in the input stream, rather than by special symbols.

The model used in the current studies exploits recurrent connections for the coding of sequential context, but does so in a way that allows a natural treatment of time (Dominey, 1998a, b). Rather than requiring that a new sequence element be processed on each time step, the model allows the experimenter to specify the durations of sequence elements and the delays between them in terms of multiple time steps. Thus, temporal structure can be specified in a realistic way. During these events with multiple time-step durations, the activity in the State network is modified in a systematic way by this passage of time, due to the dynamic flow of information in the recurrent connections as illustrated in Figure 2. Hence, the model is able to discriminate between C–V sequences that are identical in their serial order and differ only in their temporal structure, performing this task at the level of human babies as described in Nazzi et al. (1998).

Even more interesting is the fact that the extraction of both distributional serial regularities as in Saffran et al. (1996a) and temporal regularities as in Nazzi et al. (1998) is performed by the same neural architecture. Of course, in the current study these two capabilities were demonstrated in two distinct sets of simulations. Nevertheless, the architecture of the TRN makes it possible for a single physical network to represent serial and temporal structure at the same time, if the input sequences vary along both dimensions (which, again, was not the case in the two present simulations). This has been shown for visuomotor sequences where serial and temporal structures were intertwined (Dominey, 1998a, b). Indeed, serial order learning was improved in the presence of coherent or corresponding temporal structure, and impaired when the temporal structure was randomised. These sequences were formally analogous to a variant of Saffran et al.'s artificial language where syllables would also vary in duration, and the TRN was able to capture the temporal regularities at the same time as the serial ones. Thus, this series of simulations render plausible the hypothesis that a common neural architecture underlies the extraction of both serial and temporal properties

[3] Of course, there remains one fundamental level of pre-processing, the phonetic segmentation of speech. This pre-processing is actually shared by all neural network models that take speech as input (with the variant of features). Even in this respect, our simulation of rhythm processing is quite economical, since it doesn't require a full phonetic segmentation; input speech is just very broadly segmented into consonantal and vocalic intervals. Moreover, the segmentation in terms of consonants and vowels was chosen mainly for practical reasons, but it is plausible that a segmentation in terms of highs and lows in a sonority or in an energy curve would yield the same rhythmic regularities.

of speech and other sensorimotor sequences, and that this same system may perform the extraction of both types of properties together from a given type of sequence.

Furthermore, examining how the performance of the model evolves in various conditions enables to make testable predictions concerning the performance of infants. We have previously demonstrated that in a distributional regularity learning task, the TRN's performance is impaired when irrelevant temporal variation is introduced (Dominey, 1995, 1998a, b). For example, if the test sequences are generated by randomly introducing pauses in the training sequences, then performance on the test sequences is impaired (Dominey, 1998a), as observed in human adults (Stadler, 1995). Note that the random introduction of pauses will yield temporal modifications quite different from those due to speaking rate differences, for which the infant should be relatively insensitive. We thus predict that when the speech stream is contaminated with randomly introduced delays, infants should exhibit decreased performance to the same extent, which could be tested using Saffran et al.'s task with test syllables of different durations from training syllables.

We may now turn to the sensitivity to abstract structure in sequences. The lack of such a sensitivity has plagued attempts to model abstract properties of language with simple recurrent networks. For instance, Elman (1991) demonstrated that his SRN was able to learn certain aspects of the syntactic relations between nouns and verbs. However, the SRN is unable to generalise these syntactic relations to sentences made of new words lying outside the training space (see Marcus, 1998a, for a detailed argument and Marcus, 1998b, for relevant simulations). We can thus consider that the SRN doesn't provide a plausible neural implementation of this generalisation aspect of syntactic rules. The same argument applies to the numerous models (primarily feedforward) that have attempted to capture morphological rules underlying the English past tense (see for instance Rumelhart & McClelland, 1986, for such a model, and Prasada & Pinker, 1993, for criticism). Unsurprisingly, our TRN does no better in this respect than the SRN, as demonstrated by our simulation of Marcus et al. (1999). What is more interesting is to see what it takes for a neural network to be able to capture abstract relations. Dominey et al. (1998) had already proposed the addition of short-term memory and recognition modules to the TRN to account for the ability to extract abstract structure. Here, using the same architecture, the Abstract Recurrent Network, we showed that it is indeed capable of simulating infants' ability to capture abstract relationships between words of an artificial language (Marcus et al., 1999). The architecture of the ARN offers a plausible implementation of this capacity, given behavioural and neuropsychological evidence in humans that sensitivity to surface and abstract structure are dissociable

mechanisms that can possibly be related to the fronto-striatal system for the former, and to the left anterior cortex for the latter (Dominey & Georgieff, 1997; Dominey & Jeannerod, 1997; Dominey et al., 1997, 1998).

Thus, the failure of the TRN in the abstract structure learning task leads us to agree with Marcus et al. (1999) in saying that simple recurrent networks are unable to model certain crucial aspects of cognition, and language in particular. Even though, like Seidenberg (1997), we feel that the statistical properties of the input have too often been overlooked, both Marcus et al.'s experiments and our simulations show that learning cannot be reduced to the discovery of statistical regularities on the surface. This is also in agreement with Gallistel (1990), who claims that there is more to learning than the simple principles of association. Recall indeed that the changes to the TRN that allow the sensitivity to abstract structure in the ARN include the Recognition function, which is a comparator, a typically non-associationist mechanism.

This discussion of the necessary separation between the treatment of surface and abstract structure cannot be complete without addressing the following point: While the demonstration that TRN cannot solve the abstract structure task is pretty convincing, another question remains. Could the ARN also solve the first two tasks? If so, would this not open up the possibility that a single system could account for performance on all three of these structure tasks? The answer is no, for several reasons (see Dominey et al., 1998). First, and most relevant here, the ARN only represents internal repetitive structure. It will fail to discriminate surface structure differences in sequences like ABA and CDC, both of which will be represented as XYX. Even worse, in the Saffran input data, the strings have no internal repetition, and thus they will all have an equivalent representation of XYZ for the ARN, thus eliminating the possibility of a single system solution.

Finally, we were also interested in the degree to which this sensitivity to serial, temporal and abstract structure must rely on pre-exposure to the environment. Our simulation results have demonstrated that in fact this sensitivity to the relevant properties of the input can derive without learning, directly from the recurrent architecture of the recurrent State-$State_D$ network. In the simulations, learning involved the formation of associations between neural activity patterns encoding sequence context in State, and responses in Output. Learning was not required, however, to yield this context representation capability in State. The ability in the State network to generate appropriate context representations is an inherent, non-learned property of its recurrent temporal dynamics, derived from recurrent inhibitory and excitatory connections in a network of leaky integrator neurons. This shows that a great deal of sophisticated processing can be innately programmed, without the need

to pre-specify any synaptic weight, a matter of concern to Elman et al. (1996). We note that the TRN simulates both 3-day-old (Nazzi et al., 1998) and 7–8-month-old (Saffran et al., 1996a) behaviour based on its innate capabilities, and does not take into account the effects of development.

Specifically we consider the following to be innate. The architecture is as specified in the Equations 1–6, including the learning algorithms. The capability to represent serial, temporal and abstract structure as we have defined them derives directly from the architecture, in particular the characteristics of the input layer, the recurrent network, and the associative memory. This analysis of the capabilities innately implemented in the model gives us indications as to which innate abilities may be sufficient for the infants to behave as they do.[4] For the Saffran et al. experiment, babies need to be able to represent distinct syllables, and to provide this input to recurrent corticocortical networks that can then represent their conditional probabilities of occurrence in the sound sequence. For the Nazzi et al. experiment, babies need to be able to segment speech into consonants and vowels, and again their recurrent cortical network can then represent certain statistical regularities of their respective durations. In particular, no experience with the particular languages tested nor with any other language is needed prior to testing (except of course the exposure to the training conditions during the experiment), which is consistent with the very limited linguistic experience that Nazzi et al.'s 3-day-old newborns have. In the simulations we provided the initial signal in terms of either distinct syllables or consonants and vowels in the sound sequence, but the model itself performed the actual work of representing the critical relations between these events. Finally, for the Marcus et al. experiments, babies in addition need to be able to recognise recent repetitions, and to provide this input to the recurrent corticocortical context encoding network. Given these capabilities, like the babies, the models could form the appropriate behavioural associations after minimal exposure to the training material, based on intrinsic representational properties of recurrent networks. This is consistent with evidence that quite early in life these recurrent corticocortical networks are in place and functional in the frontal cortex of the baby (see Rakic, Bourgeois, & Goldman-Rakic, 1994).

---

[4] Of course we cannot demonstrate that such abilities are necessarily innate. This is just the most conservative assumption as long as we do not know any mechanism through which these abilities could be learned. For instance, we have no clue as to how an infant might learn to perceive syllables as entities, or how he or she might learn to use some neurons as a short-term memory and others as a Recognition function (no SRN has ever been shown to spontaneously modularise itself this way).

Thus, aside from the technical interest of demonstrating both serial and temporal generalisation in one recurrent network and abstract structure generalisation in a related network, these results are of particular interest from the perspective of human language acquisition. It is becoming increasingly apparent that the acquisition of lexical and syntactic knowledge in language relies on "bootstrapping" from phonological structure including distributional statistics and prosody (Christophe et al., 1997; Gleitman & Wanner, 1982; Morgan, 1986; Morgan & Demuth, 1996; see Fernald & McRoberts, 1996 for an alternative point of view). The current study characterises how sensitivity to such phonological information can be realised with minimal representational and computational capabilities, as well as with minimal experience, and can also provide some insight into the possible underlying neurophysiology (Dominey et al., 1995), awaiting more direct biological evidence. It also demonstrates that the ability to process abstract structure as observed by Marcus et al. (1999) requires additional representational capabilities, distinct from those required for processing serial and temporal structure. Clearly however, more research is needed to assess whether the ARN we proposed to solve this problem is able to learn more realistic syntactic relations than simple ABA/ABB sequences. More generally, our future research will address how this sensitivity to serial, temporal and abstract structure can contribute to acquisition and representation of syntactic structure.

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine.

Alexander, G.E., DeLong, M., & Strick, P.L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381.

Barone, P. & Joseph, J.-P. (1989). Prefrontal cortex and spatial sequencing in the macaque monkey. *Experimental Brain Research*, *78*, 447–464.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P.W., Kennedy, L.J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General*, *117*, 21–33.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Language and Speech*, *38*, 311–329.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*, 711–721.

Brent, M.R., & Cartwright, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Christiansen, M.H., Allen, J., & Seidenberg, M.S. (1998). Learning to segment speech using multiple causes: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America*, *95*, 1570–1580.

Christophe, A., Guasti, T., Nespor, M., Dupoux, E., & van Ooyen, B. (1997). Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition. *Language and Cognitive Processes*. *12*, 585–612.

Cleeremans, A., & McClelland, J.L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Church, K.W. (1987). Phonological parsing and lexical retrieval. *Cognition*, *25*, 53–69.

Cooper, W.E., & Pacia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Cutler, A., & Butterfield, S. (1990). Syllabic lengthening as a word boundary cue. *Proceedings of the 3rd Australian International Conference on Speech Science and Technology*, 324–328.

Cutler, A., & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133–142.

Dominey, P.F. (1995). Complex sensory-motor sequence learning based on recurrent state-representation and reinforcement learning. *Biological Cybernetics*, *73*, 265–274.

Dominey, P.F. (1997). An anatomically structured sensory-motor sequence learning system displays some general linguistic capacities. *Brain and Language*, *59*, 50–75.

Dominey, P.F. (1998a). Influences of temporal organization on transfer in sequence learning: Comments on Stadler (1995) and Curran and Keele (1993). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 234–248.

Dominey, P.F. (1998b). A shared system for learning serial and temporal structure of sensorimotor sequences? Evidence from simulation and human experiments. *Cognitive Brain Research*, *6*, 163–172.

Dominey, P.F., Arbib, M.A., & Joseph, J.P. (1995). A model of cortico-striatal plasticity for learning oculomotor associations and sequences. *Journal of Cognitive Neuroscience*, *7*, 311–336.

Dominey, P.F., & Boussaoud, D. (1997). Encoding behavioral context in recurrent networks of the frontostriatal system: A simulation study. *Cognitive Brain Research*, *6*, 53–65.

Dominey, P.F., & Georgieff, N. (1997). Schizophrenics learn surface but not abstract structure in a serial reaction time task. *NeuroReport*, *8*, 2877–2882.

Dominey, P.F., & Jeannerod, M. (1997). Contribution of frontostriatal function to sequence learning in Parkinson's disease: Evidence for dissociable systems. *NeuroReport*, *8*, r3–r9.

Dominey, P.F., Ventre-Dominey, J., Broussolle, E., & Jeannerod, M. (1997). Analogical transfer is effective in a serial reaction time task in Parkinson's disease: Evidence for a dissociable sequence learning mechanism. *Neuropsychologia*, *35*, 1–9.

Dominey, P.F., Lelekov, T., Ventre-Dominey, J., & Jeannerod, M. (1998). Dissociable processes for learning the surface and abstract structure sensorimotor sequences. *Journal of Cognitive Neuroscience*, *10*, 734–751.

Doy, K. (1995). Recurrent networks: Supervised learning. In M.A. Arbib (Ed.), *The handbook of brain theory and neural networks*, pp. 796–800, Cambridge MA: MIT Press.

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

Elman, J.L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–224.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fernald, A., & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pp. 365–388. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Friederici, A.D., & Wessels, J.M.I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech-perception. *Perception and Psychophysics*, *54*, 287–295.
Gallistel, C.R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books/MIT Press.
Gleitman, L.R., & Wanner, E. (1982). Language acquisition: The state of the art. In E. Wanner & L.R. Gleitman (Eds.), *Language acquisition: The state of the art*, pp. 3–48. Cambridge, UK: Cambridge University Press.
Goldman-Rakic, P.S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In V. Mountcastle (Ed.), *Handbook of physiology*, *5*, 373–417.
Harris, Z.S. (1954). Distributional structure. *Word*, *10*, 146–162.
Harris, Z.S. (1995). From phoneme to morpheme. *Language*, *31*, 190–222.
Hirsh-Pasek, K., & Golinkoff, R.M. (1996). *The origins of grammar: Evidence from early language comprehension*. Cambridge, MA: MIT Press.
Hirsh-Pasek, K., Kelmer Nelson, D.G., Jusczyk, P.W., Cassidy, K.W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*, 269–286.
Hooper, S.L. (1998). Transduction of temporal patterns by single neurons. *Nature Neuroscience*, *1*, 720–726.
Ivry, R.B., & Hazeltine, R.E. (1995). Perception and production of temporal intervals across a range of durations: Evidence for a common timing mechanism. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 3–18.
Jusczyk, P.W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
Jusczyk, P.W., & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.
Jusczyk, P.W., Luce, P.A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.
Jusczyk, P.W., Charles-Luce, J., & Luce, P. (1994). Infants' sensitivity to phonotactic patterns in their native language. *Journal of Memory & Language*, *33*, 630–645.
Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208–1221.
Kühn, R., & van Hemmen, J.L. (1992). Temporal association. In E. Domanay, J.L. van Hemmen, & K. Schulten (Eds.), *Physics of Neural Networks*, pp. 213–280. Berlin: Springer-Verlag.
Ladefoged, P. (1975). *A course in phonetics*. New York: Harcourt Brace Jovanovich.
Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*, 145–163.
Mandel, D.R., Jusczyk, P.W., & Kelmer Nelson, D.G. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, *53*, 155–180.
Marcus, G.F. (1998a). Can connectionism save constructivism? *Cognition*, *66*, 153–182.
Marcus, G.F. (1998b). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*, 243–282.
Marcus, G.F., Vijayan, S., Bandi Rao, S., & Vishton, P.M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80.
Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pp. 101–116. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*, 143–178.
Morgan, J.L. (1986). *From simple input to complex grammar*. Cambridge, MA: MIT Press.

Morgan, J.L. (1994). Converging measures of speech segmentation in preverbal infants. *Infant Behavior and Development*, *17*, 389–403.
Morgan, J.L. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language*, *35*, 666–688.
Morgan, J.L., & Demuth, K. (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
Morgan, J.L., & Saffran, J.L. (1995). Emerging integration of sequential and supra-segmental information in preverbal speech segmentation. *Child Development*, *66*, 911–936.
Nakatani, L.H., & Schaffer, J.A. (1978). Hearing words without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America*, *63*, 234–245.
Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology, Human Perception and Performance*, *24*, 1–11.
Nespor, M., Guasti, T., & Christophe, A. (1996). Selecting word order: The rhythmic activation principle. In U. Kleinhens (Ed.), *Interfaces in phonology*, pp. 1–26. Berlin: Akademic, Verlag.
Nissen, M.J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.
Pearlmutter, B.A. (1995). Gradient calculation for dynamic recurrent neural networks: A survey. *IEEE Transactions on Neural Networks*, *6*, 1212–1228.
Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.
Pike, K.L. (1945). *The intonation of American English*. Michigan: Ann Arbor.
Pineda, F.J. (1989). Recurrent backpropagation and the dynamical approach to adaptive neural computation. *Neural Computation*, *1*, 161–172.
Prasada, S., & Pinker, S. (1993). Similarity-based and rule-based generalizations in inflectional morphology. *Language and Cognitive Processes*, *8*, 1–56.
Rakic, P., Bourgeois, J.-P., & Goldman-Rakic, P.S. (1994). Synaptic development of the cerebral cortex: Implications for learning, memory and mental illness. In J. van Pelt, M.A. Corner, H.B.M. Uylings, & F.H. Lopes da Silva (Eds.), *Progress in Brain Research*, pp. 227–243.
Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, *105*, 512–521.
Ramus, F., Nespor, M., & Mehler, J. (in press). Correlates of linguistic rhythm in the speech signal. *Cognition*.
Redington, M., Chater, N., & Finsh, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425–469.
Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tenses of English verbs. In J.L. McClelland, D.E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2: *Psychological and Biological Models*, pp. 216–271). Cambridge, MA: MIT Press.
Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
Seidenberg, M.S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*, 1599–1603.
Shi, R., Morgan, J.L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, *25*, 169–201.

Stadler, M.A. (1995). The role of attention in implicit learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 674–685.

Van Ooyen, B., Bertoncini, J., Sansavini, A., & Mehler, J. (1997). Do weak syllables count for newborns? *Journal of the Acoustical Society of America*, *102*(6), 3735–3741.

Walsh, J.P., & Dunia, R. (1993). Synaptic activation of N-methyl-D-aspertate receptors induces short-term potentiation at excitatory synapses in the striatum of the rat. *Neuroscience*, *57*, 241–248.

Weitzenfeld, A. (1991). NSL Neural Simulation Language, v 2.1, Center for Neural Engineering, University of Southern California Technical Report TR91-5.

Werbos, P.J. (1995). Backpropagation: Basics and new developments. In M.A. Arbib (Ed.), *The handbook of brain theory and neural networks*, pp. 134–139. Cambridge MA: MIT Press.