TRACX2: A connectionist autoencoder using graded chunks to model infant visual statistical learning

Denis Mareschal[1] & Robert M. French[2]

[1]Centre for Cognition, Computation and Modelling
Centre for Brain and Cognitive Development
Birkbeck University of London
UK

[2] Laboratoire d'Etude de l'Apprentissage et du Développement
CNRS UMR 5022
Univeristé de Bourgogne-Franche-Comté
Dijon, France

Acknowledgements: This work was funded by Economic and Social Research Council Grant RES-360-25-056 to the first author and an grant from the Agence Nationale de la Recherche Scientifique (ANR), ANR-14-CE28-0017, to the second author. DM is further funded by a Royal Society Wolfson Research Merit Award.

Address all correspondence to d.mareschal@bbk.ac.uk or robert.french@u-bourgogne.fr. All code used for the simulations reported here can be obtained from Robert French.

Author Contributions: DM & RF developed the ideas, RF ran the simulations, DM & RF wrote the article.

Abstract

Even newborn infants are able to extract structure from a stream of sensory inputs and yet, how this is achieved remains largely a mystery. We present a connectionist autoencoder model, TRACX2, that learns to extract sequence structure by gradually constructing chunks, storing these chunks in a distributed manner across its synaptic weights, and recognizing these chunks when they re-occur in the input stream. Chunks are graded rather then all-or-none in nature. As chunks are learnt their component parts become more and more tightly bound together. TRACX2 successfully models the data from five experiments from the infant visual statistical-learning literature, including tasks involving forward and backward transitional probabilities, low-salience embedded chunk items, part-sequences, and illusory items. The model also captures performance differences across ages through the tuning of a single learning rate parameter. These results suggest that infant statistical learning is underpinned by the same domain general learning mechanism that operates in auditory statistical learning and, potentially, in adult artificial grammar learning.

1. Introduction

We live in a world in which events evolve over time. Consequently, our senses are bombarded with information that varies sequentially through time. One the greatest challenges for cognition is to find structure within this stream of experiences ([1]; [2]). Even newborn infants are able to do this ([3]; [4]), and yet, how this is achieved remains largely a mystery.

Two possibilities have been suggested (see [5], [6], Theissen (this issue) for detailed discussions). The first, often referred to as *statistical learning*, involves learning the frequencies and transitional probabilities (TPs) of an input signal to construct an internal representation of the regularity boundaries between elements encountered (e.g., [7]; [8]). The second possibility, often referred to as *chunking*, suggests that elements that co-occur are simply grouped together – or chunked – into single units, stored in memory and recalled when necessary [9]. Over time, these chunks can themselves be grouped into super-chunks or super-units. According to this view behaviour is determined by the recognition of these chunks stored in memory and associated with particular responses (e.g., [5]; [9]; [10]). What distinguishes these accounts is that the former argues that it is the probabilistic structure of the input sequence that is represented and stored (e.g., TPs), whereas the later argues that specific co-occurring elements are stored, rather than the overarching statistical structure. Ample evidence in support of both of these views has been reported in the literature.

We will argue that both transitional probability learning (statistical learning) and chunking co-exist in one system implementing a single learning mechanism, which can transition smoothly between two apparently distinct modes of behaviour. The appearance of two modes of learning is an illusion because only a single mechanism underlies sequential learning; namely, Hebbian-style learning in a partially recurrent distributed neural network. Such a system encodes exemplars (typical of chunking mechanisms) while drawing on co-occurrence statistics (typical of statistical learning models). An important corollary of this approach is that *chunks are graded* in nature rather than all-or-none. Moreover, interference effects between chunks will follow a similarity gradient typical of other distributed neural network memory systems.

Chunks were historically thought of as all-or-nothing items ([9], [11]). However, recent work (for example on the gradedness of the morphological features of compound word, [12], [13]) shows that this is not the case. When we encounter the words "smartphone", "carwash", or "petshop", we still clearly hear the component words. We hear them less in words like "sunburn" and "heartbeat". We hear them hardly at all in "automobile." How long did it take for people to stop hearing "auto" and "mobile" when they heard or read the word "automobile"? Like "automobile", it is likely that in a few years the current generation will no longer hear "smart" and "phone" when they hear the word "smartphone". This simple observation involving the graded nature and gradual lexicalisation of chunks is at the heart of the chunking mechanism in TRACX2.

In TRACX [14] we showed that a connectionist autoencoder, augmented with conditional recurrence, could extract chunks from a stream of sequentially presented inputs. TRACX had two banks of input units, which it learnt to autoencode onto two banks of identical output units. Sequential information was encoded by presenting successive elements of the sequence, first on the right input bank, then on the left input bank on the next time step. Thus, the sequence of inputs was presented in a successive series of right-to-left inputs, with learning occurring at each time step. However, if the output autoencoding error was below some pre-set threshold value (indicating successful recognition of the current pair of input elements), then, on the next time step, instead of the input to the right input bank being transferred to the left input bank, the *hidden unit representation* was put into the left input bank. The next item in the sequence was, as always, put into the right input bank. Weights were updated and the input sequence would then proceed as before. The result of this was that TRACX learnt to form chunks of elements that it recognised as co-occurring (see [14] for full details). TRACX successfully captured a broad range of data from the adult and infant auditory statistical learning literature (e.g., [15]; [16]; [17]; [18]). Moreover, it outperformed existing models of both chunking, notably, PARSER ([19]; [10]) and statistical learning (SRNs, [20]). Finally, the model was able to scale up to more realistic linguistic *corpora*, in particular, the Brent & Cartwright [21] data.

In the present article, we introduce TRACX2, an updated version of TRACX, which removes the use of an all-or-nothing error threshold that determines whether or not the items on input are to be chunked. This effectively removes a mechanism

— a conditional jump (i.e. an *if-then-else*) statement — that is not natural to neural network computation. In TRACX2, the contribution of the hidden-unit activation vector to the left bank of input units is graded and depends on the level of learning already achieved. We then use TRACX2 to model a total of seven experiments, two classic experiments from the infant auditory statistical learning literature that we previously modelled with TRACX ([14]) and five from the infant visual statistical learning literature. Visual statistical learning paradigms involve showing infants sequences of looming coloured shapes with varying degrees of statistical regularity embedded in the sequences. It was first developed as a visual analogue of the auditory statistical learning experiments ([22]) and has yet to be captured by any modelling paradigm.

In summary, the aim of this article is: (1) to describe the TRACX2 architecture, (2) to model a range of representative phenomena characteristic of infant visual statistical learning with the TRACX2 architecture and, as a result, (3) to demonstrate that behaviours typically taken as evidence of either a chunking or statistical learning mechanisms can be accounted for by a single learning mechanism.

2. **The TRACX2 Architecture**

TRACX2 was initially introduced by French and Cottrell [23]. This recurrent connectionist model consists of an autoencoder with two identical banks of inputs units, two identical banks of output units (each of which is the same size as each of the banks of input units), and a bank of hidden units with the same dimensions as

one of the input/output unit banks (Figure 1). In the current implementation, the model is trained using the backpropagation algorithm.

========== Insert Figure 1 about here =======

The key to understanding TRACX2 is to understand the flow of information within the network. Over successive time steps, the sequence of information is presented item-by-item into the right-hand bank (RHS) of input units. The left-hand bank (LHS) of input units is filled with a blend of the right-hand input and the hidden unit activations at the previous time step. This mixture is determined by Equation 1:

$$LHS_{t+1} = (1- \tanh(\alpha\Delta_t))*Hiddens_t + (\tanh(\alpha\Delta_t))*RHS_t \qquad \text{(Eq 1)}$$

where Δ_t is the absolute value of the maximum error across all output nodes at time *t,* LHS_t is the activation across the left-hand bank of input nodes, $Hiddens_t$ are the hidden-unit activations at time *t*, and RHS_t is the activation across the right-hand bank of input nodes. Finally, α determines the weight of the contribution of the internal representation at time *t* to the left-hand bank of inputs at time *t+1*. Unless otherwise stated, for all simulations in this paper we have set α to 1. If at time t, Δ_t is small, this means that the network has learnt that the items on input are frequently together (otherwise Δ_t could not be small). The contribution to the left-hand bank of input units at time t+1 of the hidden-unit activations, which constitute the network's internal representation of the two items on input at time t, is, therefore, relatively large and the contribution from the right-hand inputs will be relatively small. Conversely, if Δ_t is large, meaning that the items on input have not been seen

together often, the hidden-layer's contribution at time *t+1* to the left-hand input bank will be relatively small and that from the right-hand inputs will be relatively large. Finally, at each time step, the weights are updated to minimise the output autoencoder error.

In layman's terms, this means that as you experience items (visual, auditory, tactile) together over and over again, these items become bound to each other more and more strongly into a chunk. At first, a chunk is weak (e.g., "smartphone"), but if it is encountered frequently, it gradually develops into a tightly bound chunk in which we no longer perceive its component parts.

3. **Modelling infant statistical learning**

In this section we report on a total of seven different simulations using TRACX2 of infant statistical learning behaviour, two from classic studies in the auditory domain ([11]; [12]), and the remainder from the visual domain. All weights were initially randomised between -1 and 1. The Δ value determining the amount of new input vs. hidden unit representation presented at input was determined by the maximum absolute error over all output units. So, for example, if Desired Output = [0.1, 0.5, 0.4] and Actual Output = [0.3, 0.9, 0.3], then, the absolute difference between the two is [0.2, 0.4, 0.1], and the max-abs-diff over the three units is Δ=0.4. Note that for updating weights in the network, we used the standard summed-squared-error typical of Backpropagation networks. There was no momentum term, but a Fahlman offset of 0.01 was used. We used a Tanh squashing function to determine the hidden and output unit activations. Finally, all simulations are averages over 30 runs.

We used η (the learning rate) as a proxy for development, with η set to 0.0005 for newborns, 0.0015 for 2 month-olds, 0.0025 for 5-month-olds, and 0.005 for 8-month-olds. There was a bias node on the input and hidden layers and momentum was always set to 0. The key developmental hypothesis here is that, with increasing age, infants are progressively better at taking up information from an identical environment. This is consistent with the well-established finding that the average rate of habituation increases with increasing age during infancy (e.g., [24]; [25] ; [26]). Finally, as has been used repeatedly elsewhere, we take network output error as a proxy for looking time in the infant ([27] ; [28] ; [29] ; [30]; [31]; [32]; [26]). The idea here is that the amount of output error correlates with the number of cycles required to reduce the initial error, which corresponds to the amount of time or attention that the model will direct to a particular stimulus.

The first two simulations are replications by TRACX2 of results reported in French et al. [14] and French and Cottrell [23]. We show that TRACX2 captures the key phenomena in auditory statistical learning (i.e., [11] and [12]). Next, we model the seminal Kirkham et al. [22] visual statistical learning experiment demonstrating that age-related effects in the efficacy of learning can be accounted for by a simple and plausible parameter manipulation in TRACX2. We then show that TRACX2 can capture statistical learning in newborns, as well as their dependency on the complexity of the information stream ([4]). Next, we show that TRACX2 captures the processing of backward transitional probabilities ([18]; [33]) in much the same way as 8-month-olds ([34]). Finally, we show that, like 8-month-olds ([35]; [36]),

TRACX2 forms illusory conjunctions, normally taken as evidence of a statistical (TP) learning mechanism, but also shows decreased salience of embedded chunk items, normally taken as evidence of a chunking mechanism. It, therefore, reconciles two apparently paradoxical infant behaviours within a single common mechanism.

3.1 Auditory statistical learning

Saffran, Aslin & Newport [15] is a seminal paper on infant syllable-sequence segmentation. Six different words were used, each with 3 distinct syllables from a 12-syllable alphabet. A random sequence of 90 of these words (270 syllables) with no immediate repeats or pauses between words was presented twice to 8-month-old infants. After this familiarisation period, the infants heard a word from the familiarisation sequence and a partword from that sequence. A head-turn preference procedure was used to show that infants had a novelty preference for partwords. The conclusion of the authors was that the infants had learnt words better than partwords. We simulated this experiment with TRACX2 and a typical SRN[1] using the same number of words drawn from a 12-syllable alphabet. The familiarisation sequence was the same length as the one that the infants heard. Both models learnt words better than partwords. Note also that, although the SRN performance seems to deviate more from infant performance than that of TRACX2, we did not carry out a systematic search for the optimal SRN parameters, so it may be possible that better SRN performance could be achieved with different parameters.

[1] A 24-12-24 architecture was used with a learning rate of 0.01 and momentum of 0.9 with a Fahlman offset of 0.1. Bipolar (i.e., -1, 1) orthogonal encodings localist encodings were used for each of the 12 syllables.

========= Insert Figure 2 about here ========

However, in Saffran et al. [15] there was a confound – namely, words were heard three times as often as partwords. Aslin, Saffran & Newport [16] then designed an experiment that removed the unbalanced frequency of words and partwords. There were now four 3-syllable words, two of which occurred twice as often in the familiarisation sequence as the other two. Thus, the partwords spanning the two high-frequency words would have the same overall frequency in the familiarisation sequence as the low-frequency words. The same head-turn preference procedure showed, again, that infants had a novelty preference for partwords. These authors' conclusion was that the infants had learnt words better than partwords. Once again, we designed a set of words exactly like those used in [16]. The length of the familiarisation sequence was also identical to that used in [16]. Figure 2 shows the performance of 8-month-old infants, TRACX2 and a simple recurrent network (SRN) on words and partwords from this sequence.

We can also use these data to illustrate the role of the α parameter in TRACX2. This parameter controls the extent to which hidden unit representations are incorporated into the left-side input representations. If α is large, then Δ (error) has to be extremely small before the hidden layer begins to contribute to the left-hand-side input. Under these circumstances, the network will find it very hard, if not impossible, to form chunks larger than two successive items that can be encoded across the two banks of input units. In other words, if α is too large, there will be little or no internal (i.e., hidden-unit) contribution to the left-hand-side input units. On the other hand, if α is too small, the contribution from the hidden layer to the left-hand bank of units will always be

significant, whether or not the previous two items on input had been seen together frequently by the network. This is largely irrelevant in many of the infant visual statistical learning experiments because "words" tend to consist of only two images. However, [16] and [15] use three-element words. As can be seen in Figure 2b, if α is too small or too big, then TRACX2 is unable to chunk 3-elements into a single word, and is therefore unable to differentiate 3-element words from part-words. For all of the simulations reported in this article, we set α to 1, which allowed good chunking.

3.2 Visual statistical learning

Kirkham et al [22] developed a visual analogue of the auditory statistical learning tasks initially developed by Saffran, Aslin and colleagues [15]. Instead of listening to unbroken streams of sounds, infants were shown continuous streams of looming colourful shapes in which successive visual elements within a "visual word" were deterministic, but transitions between words were probabilistic (see Figure 3). Infants at three different ages were first familiarised to this stream of shapes, then presented with either a stream made up of the same shapes but with random transitions between all elements, or a stream made up of the identical visual words as during habituation. Kirkham et al. [18] found that infants from 2 months of age subsequently looked longer at the random sequence than the structured sequence (even though the individual elements are identical between streams) suggesting that the infants had learnt the statistical structure (TPs) of the training sequence.

========== insert Figure 3 about here =======

We modelled this experiment by training the model with a sequence of inputs containing the identical probability structure to that used to train infants. The training sequence was identical in length to that used by Kirkham [22]. The transitional probability within a visual word was p=1.0, and between visual words p=.33. Shapes were coded using localist, bipolar (i.e., -1, 1) orthogonal encodings in order minimise effects due to input similarity. As in the Aslin et al. [16] and Saffran et al. [15] simulations, the RHS and LHS input vectors were comprised of 12 units. Network performance was evaluated by averaging output error over all three of the possible two-image ”visual words" in the sequence. This was then compared to the average output error for a set of three randomly selected two-image “visual non-words” that were neither words nor part-words, and consequently, occurred nowhere in the training sequence. This is analogous to the word/non-word testing procedure used in auditory statistical learning studies (e.g., Saffran et al., [15]), and completely equivalent to testing the networks with a structured sequence (from which they would have extracted visual words) and a fully random sequence (in which no previous words or part-words exist). The model, like infants at all ages, looked longer at the randomised sequence than the structured sequence (Figure 3a).

3.3 Visual statistical learning in newborns

Bulf, Johnson, and Vilenza [4] asked whether the sequence-learning abilities demonstrated by Kirkham et al [22] were present from birth. They tested newborns (within 3 days of birth) on black and white sequences of streaming shapes. In their “High

Demand Condition", the sequence had the same statistical structure as in Kirkham et al [18]. That is, the sequences were made up of 3 visual words, each made up of two shapes with a constant transition probability of 1.0 defining the word, and transitional probabilities of .33 between words. They also introduced a "Low Demand Condition" in which the sequences were made up of only two words (each consisting of two shapes with internal transition probabilities of 1.0) leading to transition probabilities at word boundaries of 0.5 (instead of the .33 previously used). The reasoning here was that newborns had more limited information processing abilities and may therefore struggle with a more complex sequence, already proving to be a challenge for 2-month-olds.

==========Insert Figure 4 about here =====

Again, we modelled this study using TRACX2, in the same way as above, but by (1) reducing the learning rate to 0.0005, and (2) creating both high-demand and low-demand sequences. In the low-demand condition (LDC), there were two pairs of images, each made up of two different images (i.e., a total of 4 separate images). In the high-demand condition (HDC) there were three pairs of images, each made up of two different images (i.e., a total of 6 separate images). In the simulation for both the high-demand and low-demand conditions, TRACX2 saw sequences of 120 words. Statistics were averaged over 30 runs of the program, with each run consisting of 10 simulated subjects. Figure 4 shows both the infant data and the model results. As with the infants, TRACX2 did not discriminate between the structured training sequence and the random sequence in the high demand condition (with the lower learning rate) but did discriminate between the

two sequences in the low demand condition.

3.4 *Learning backward transitional probabilities*

Tummeltshammer and colleagues [34] explored whether 8-month-olds could utilise backward transitional probabilities, as well as forward transitional probabilities, to segment the looming shape sequences. Backward transitional probabilities occur when there is a high probability that an item is *preceded* by something rather than the other way around ([18]; [33]). While the original TRACX model was able to capture the infant and adult data related to the processing of backward transitional probabilities in auditory sequences, SRNs were not able to do so ([10]). This is, therefore, an important test of the underlying learning architecture. For the simulations we used a sequence containing 48 items taken from Table 1 of [30]. In the actual experiment with infants, this sequence was repeated only 3 times, but for our simulation we found that this did not produce sufficient learning and we used a training sequence that was produced by repeating this sequence 25 times. The learning rate was set at 0.005. Figure 5 shows that both 8-month-olds and TRACX2 are able to segment sequences involving predictable backward transitional probabilities as well as sequences containing forward transitional probabilities.

======== Insert Figure 5 about here =====

3.5 Learning embedded and illusory items.

An embedded item is a group of syllables that occurs within a word, but never occurs independently (e.g., "eleph", as in "elephant"; Thiessen et al., 2013).

Statistical (TP) learning accounts predict that, because learners represent the statistical relations between all pairs of adjacent elements, distinguishing components embedded in longer word should improve with greater exposure to the word. In contrast, chunking models predict that as learners become familiar with a word, they should become *less* able to distinguish sub-components embedded in that word ([17]). Thus, the recognition of illusory items and embedded items provide critical tests of the statistical learning and chunking accounts of sequence processing.

Illusory items – are pairs or triplets of elements that have never been encountered, but which have the same statistical structure (e.g., TPs) as other pairs or triplets that have been previously encountered (cf. [35]). For example, if *tazepi*, *mizeru*, and *tanoru,* are words presented in a speech stream, with TPs of p=.50 between the successive syllables in these words, then *tazeru* would be a statistically matched illusory word because the TPs between the successive syllables in this new word match the TPs encountered previously. Statistical (TP) learning mechanisms would unable to distinguish between real and illusionary words because they are statistically equivalent. In contrast, chunking mechanisms will fail to recognise the new illusory word precisely because it has never been encountered before and is therefore not stored in memory.

Fortunately, Slone & Johnson ([36]; [37]) have investigated whether infants' learning mechanisms would lead to the reduced salience of embedded items or to the emergence of illusory chunks, as a means of testing whether chunking or statistical learning (TPs) underpins infant visual sequential statistical learning. To do this, they

presented 8-months-olds with sequences structured as depicted in Figure 6a. Infants in the “Embedded Pair Experiment” did not differentiate embedded pairs from part-pairs that crossed word boundaries, but both were differentiated from the word pairs. Infants in the “Illusory Item Experiment” did not differentiate the illusory triplets from the part triplets, but both were differentiated from the actual triplets. This is perplexing because the former result suggests that infants utilise chunking, whereas the latter results suggests that they engage in statistical (TP) learning.

TRACX2 captures both of these results equally well. Recall that the model is designed to produce the smallest error on the best learnt patterns. If we consider output error to be a measure of visual attention (the higher the error, the longer the infant attends to that item), then we can say that TRACX2 is designed to orient to novel test patterns most (i.e., shows a novelty preference). A familiarity preferences is the inverse of a novelty preferences. This means that the *smaller* the error for an item, the longer the infant looks at that item. Thus, to model familiarity preferences we subtract the error on output from the maximum possible error and call this "Inverse Error" (Figure 6b). So, when modelling a *familiarity* preference, the greater TRACX2's Inverse Error, the longer the infant looking time is.

Such shifts in orienting behaviour are common in infant visual orienting, and have been related to the complexity of the stimuli and the depth of processing [38]; [39]; see also [40], for a process account of the familiarity-to-novelty shift in a neural network model of habituation). In sum, TRACX2 captures both the reduced salience of embedded chunk items and the appearance of illusory conjunctions within a single mechanism, thereby reconciling apparently paradoxical infant behaviours.

Discussion

TRACX2 ([19]) is an updated version the TRACX architecture ([14]). As in the original architecture, TRACX2 is a memory-based chunk-extraction architecture. Because it is implemented as a recurrent connectionist autoencoder in the Recursive Auto-Associative Memory (RAAM) family of architectures ([41]; [42]), it is also naturally sensitive to distributions statistics in its environment. In TRACX2, we replace the arbitrary all-or-none chunk-learning decision mechanism with a smooth blending parameter. TRACX2 learns chunks in a graded fashion as a function of its familiarity with the material presented. An implication of this is that chunks are no longer to be thought of as "all-or-none" entities. Rather, there is a continuum of chunks whose elements are bound together more or less strongly. Finally, unlike some other chunking systems such as PARSER, TRACX2 also synthesises information across prior exemplars stored in memory.

TRACX2 was used to model a representative range of infant visual statistical learning phenomena. No previous mechanistic model of these infant behaviours exist (though see [43] for a Bayesian description of adult performance on visual spatial statistical learning]. As with the auditory learning behaviours, TRACX2 captures the apparent utilisation of forward and backward transitional probabilities, the diminishing sensitivity to embedded items in the sequence, and the emergence of illusory words. However, it is important to understand that TRACX2 is not simply internalising the overall statistical structure of the sequence, but encoding, remembering and recognising

previously seen chunks of information. This is a fundamentally different account of infant behaviours than has previously been proposed (see [49]), and fits better with the recent suggestion that much of infant statistical learning can be accounted for by a memory-based chunking model ([50].

TRACX2 can use frequency of occurrence or transitional probabilities equally well and fluidly to learn a task (as is the case with 8-month-olds; [51]). This would suggest that categorizing learning either as statistical or memory-based is a false dichotomy. Both classes of behaviours can emerge from a single mechanism. The different modes of behaviour appear depending on the constraints of the task, the level of learning and the level of prior experience. Moreover, the idea that infant looking time is determined by the recognition of regularly re-occurring items (chunks or individual items) is consistent with the recent evidence suggesting that local redundancy in the sequences is the prime predictor of looking away in infant visual statistical learning experiments ([52]).

TRACX2 also suggests that there are no specialised mechanisms in the brain dedicated to sequence learning. Instead, sequence processing emerges from the application of fairly ubiquitous associative mechanisms, coupled with graded top-down re-entrant processing. Although there may be differences in the speed and richness of encoding across modalities, there is nothing intrinsically different in the way TRACX2 processes visual or auditory information. This suggests than any modality-specific empirical differences observed can be attributed to encoding differences rather than core sequence-processing differences (see Arciuli, this issue, for further discussion of the implications of differences in encoding stimuli for the understanding of individual

differences on statistical learning tasks).

In conclusion, we believe that chunking cannot be viewed as an all-or-nothing phenomenon, that learning from transitional probabilities should not be held in opposition to learning chunks. Instead, graded chunks emerge gradually precisely because of the TPs present in the input. Chunks are learnt and, over the course of being learnt, their component parts become more and more tightly bound together. This is a fundamental principle of TRACX2. The results of the present paper suggest that infant sequential statistical learning is underpinned by the same domain general learning mechanism that operates in auditory statistical learning and, potentially, also in adult artificial grammar learning. TRACX2, therefore, offers a parsimonious account of how infants find structure in time.

References

[1] Elman, J.L. (1990) Finding structure in time. *Cognitive Science*, *14*, 179-211. doi:10.1016/0364-0213(90)90002-E

[2] Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences, 10*, 233-238. doi: 10.1016/j.tics.2006.03.006

[3] Teinonen, T., Fellman, V., Näätänen, R., Alku, P., and Huotilainen, M. (2009). Statistical language learning in neonates revealed by eventrelated brain potentials. *BMC Neurosci*. 10:21. doi:10.1186/1471-2202-10-21

[4]Bulf, H., Johnson, S. P., & Valenza, E. (2011) Visual statistical learning in the newborn infant. *Cognition, 121*, 127-132. doi:10.1016/j.cognition.2011.06.010

[5] Thiessen, E. D., Kronstein, A. T., and Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*, 792. DOI: 10.1037/a0030801

[6] Pothos, E.M. (2007) Theories of artificial grammar learning *Psychological Bulletin, 133(2)* 227-244. DOI: 10.1037/0033-2909.133.2.227

[7] Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855–863. doi:10.1016/S0022-5371(67)80149-X

[8] Cleeremans, A. (1993). *Mechanisms of implicit learning*. Cambridge, MA: The MIT Press. doi:10.1007/BF00114843

[9] Gobet, F., Lane, P.C.R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I, & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*, 236-243.

[10] Perruchet, P. and Vinter, A. (2002) The Self-Organizing Consciousness. *Behavioral and Brain Sciences*, *25*, 297- 330. DOI: 10.1017/S0140525X02550068

[11] Newell, A. (1990) *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

[12] Gonnerman, L. M, Seidenberg, M. S., & Andersen, E. S. (2007) Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General, 136*, 323-345.

[13] Hay, J. B. & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences, 9*, 342-348.

[14] French, R. M., Addyman, C. & Mareschal, D. (2011) TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review, 118, 614-636*.

[15] Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996) Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928. doi:10.1126/science.274.5294.1926

[16] Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321-324. doi: 10.1111/1467-9280.00063

[17] Giroux, I. and Rey, A. (2009) Lexical and Sublexical Units in Speech Perception *Cognitive Science*, *33(2),* 260-272. doi: 10.1111/j.1551-6709.2009.01012.x.

[18] Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition, 36*, 1299-1305. doi:

10.3758/MC.36.7.1299

[19] Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246-263. doi: 10.1006/jmla.1998.2576

[20] Cleeremans, A. and McClelland, J. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7*, 161-193. DOI 10.1023/A:1022647012398

[21] Brent, M. and Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61(1),* 93-125. DOI: 10.1016/S0010-0277(96)00719-6

[22] Kirkham, N.Z., Slemmer, J.A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. *Cognition, 83,* B35-B42. DOI: 10.1016/S0010-0277(02)00004-5

[23] French, R. M. and Cottrell, G. (2014). TRACX 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction. In P. Bello, M. Guarini, M. McShane and B. Scassellati (Eds.), *Proceedings of the Thirty-sixth Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2016-2221.

[24] Bornstein, M. H., Pecheux, M.G, Lecuyer, R. (1988) Visual habituation in human infants: development and rearing circumstances. *Psychological Research, 50*, 130-133.

[25] Colombo, J. & Mitchell, D. W. (2009) Infant visual habituation. *Neurobiology of Learning & Memory, 92,* 225–234. doi:10.1016/j.nlm.2008.06.002

[26] Westermann, G. & Mareschal, D. (2013) From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B, 369*: 201220391 DOI: 10.1098/rstb.2012.0391

[27] Mareschal, D. & French, R. M. (2000) Mechanisms of categorisation in infancy. *Infancy, 1,* 59-76. DOI: 10.1207/S15327078IN0101_06

[28] Mareschal, D., French, R. M. & Quinn, P. (2000) A connectionist account of asymmetric category learning in infancy. *Developmental Psychology, 36*, 635-645. DOI: 10.1037//0012-1649.36.5.635

[29] Schafer, G. & Mareschal, D. (2001) Modeling infant speech sound discrimination using simple associative networks. *Infancy, 2,* 7-28. DOI: 10.1207/S15327078IN0201_2

[30] Mareschal, D., Quinn, P. C. & French, R. M. (2002) Asymmetric interference in 3- to 4-month-olds' sequential category learning. *Cognitive Science, 26*, 377-389. DOI: 10.1207/s15516709cog2603_8

[31] Mareschal, D. & Johnson, S. P. (2002) Learning to perceive object unity: A connectionist account. *Developmental Science, 5*, 151-172. DOI: 10.1111/1467-7687.t01-1-00217

[32] French, R. M., Mareschal, D., Mermillod, M. & Quinn, P. (2004) The role of bottom-up processing in perceptual categorization by 3- to 4-month old infants: Simulations and data. *Journal of Experimental Psychology: General, 133,*_382-397. DOI: 10.1037/0096-3445.133.3.382

[33] Pelucchi, B., Hay, J.F., & Saffran, J.R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 113*, 244-247. doi:

10.1016/j.cognition.2009.07.011

[34] Tummeltshammer, K., Amso, D., French, R. M. & Kirkham, N. (in press) Across space and time: Infants learn from backward and forward visual statistics. *Developmental Science*.

[35] Endress, A. and Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60*, 351-367. DOI: 10.1016/j.jml.2008.10.003

[36] Slone, L. K. & Johnson, S. P. (2015) Statistical and chunking processes in infants' and adults' visual statistical learning. Poster presented and the *Biannual Conference of the Society for Research in Child Development*. April 2015, Philadelphia, USA.

[37] Slone, L. K. & Johnson, S. P. (manuscript under review) When learning goes beyond statistics: Infants represent visual sequences in terms of chunks.

[38] Roder, B.J., Bushnell, E.W., & Sassville, A.M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing *Infancy, 1*, 491-507. DOI:10.1207/S15327078IN0104_9

[39] Hunter, M. A. & Ames, E. W. (1988) A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research,5*, 69-95.

[40] Sirois, S. & Mareschal, D. (2004). An interacting systems model of infant habituation. *Journal of Cognitive Neuroscience* 16, 1352-1362.

[41] Pollack, J. (1989) Implications of Recursive Distributed Representations. In David S. Touretzky (ed.) *Advances in Neural Information Processing Systems I* (pp. 527-536). Morgan Kaufmann, Los Gatos, CA.

[42] Pollack, J. (1990) Recursive Distributed Representations *Artificial Intelligence*, *46*,

77-105.

[43] Orban, G., Fiser, J., Aslin, R N., & Langyel, M. (2008), Bayesian learning of visual chunks by human observers. *Procedings of the National Academy of Science of the U. S. A., 105,* 2745-2750.

[49] Krogh, L., Vlach, H. A & Johnson, S. P (2013) Statistical learning across development: flexible yet constrained. *Frontiers in Psychology, 3*, art. 598. doi: 10.3389/fpsyg.2012.00598

[50] Thiessen, E. D. & Pavlik, P. I. (2013) iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science, 37*, 310-343.

[51] Marcovitch, S. & Lewkowicz (2009) Sequence learning in infancy: the independent contributions of conditional probability and pair frequency information. *Developmental Science, 12*, 1020-1025.

[52] Addyman, C. & Mareschal, D. (2013) Local redundancy governs infants' spontaneous orienting to visual-temporal sequences . *Child Development*, *84*, 1137-1144.

Figure Captions

Figure 1. Architecture and information flow in TRACX2. In all simulations reported in this paper, $\alpha = 1$, unless otherwise stated. When Δ is large (items not recognized as having been seen together before on input), almost all contribution to LHS comes from RHS. When Δ is small (items recognized as having been seen together before on input), almost all contribution to LHS comes from the Hidden layer (Hid).

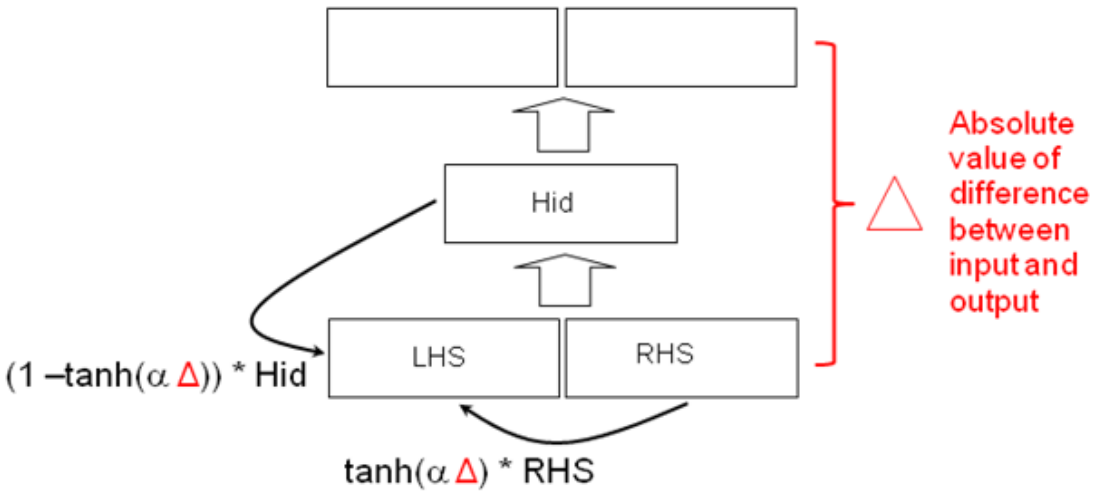
Figure 2. (a) Proportions better listening to the part words than words in infants, TRACX2, and a standard SRN. (b) Effect of varying the Tanh weighting parameter, α, in learning three-syllable words. Error on output initially falls, reaches a minimum, and then rises again.

Figure 3. (a) Illustration of visual sequences used to test infants (after Addyman & Mareschal, 2013). (b) Left-hand panel: Infant performance reported in [18] and, right-hand panel: TRACX2 performance with the familiar structured and novel non-structured sequences. (Error is the maximum error of the network over all output units; SEM error bars.)

Figure 4. Newborn performance as reported in [4] in left panel and TRACX2 performance in right panel for familiar structured and novel non-structured sequence.

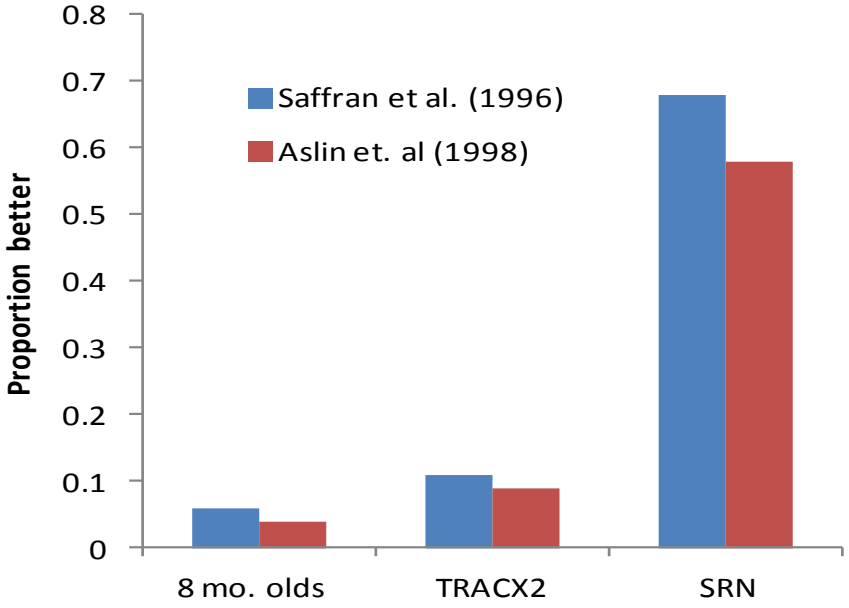
Figure 5. Infant and TRACX 2 performance when trained on sequences with either forward to backward transitional probabilities. Left panel reproduced from [30] with permission.

Figure 6. (a) Familiarisation and testing items for embedded pairs (left panel) and illusory items (right panel) (after [31] [32]). (b) Infant data (left-hand side of figure) and TRACX2 performance (right-hand side, SEM error bars). Top row: familiarity preference, Experiment 1; Bottom row: Novelty preference, Experiment 2. Figure permission pending.

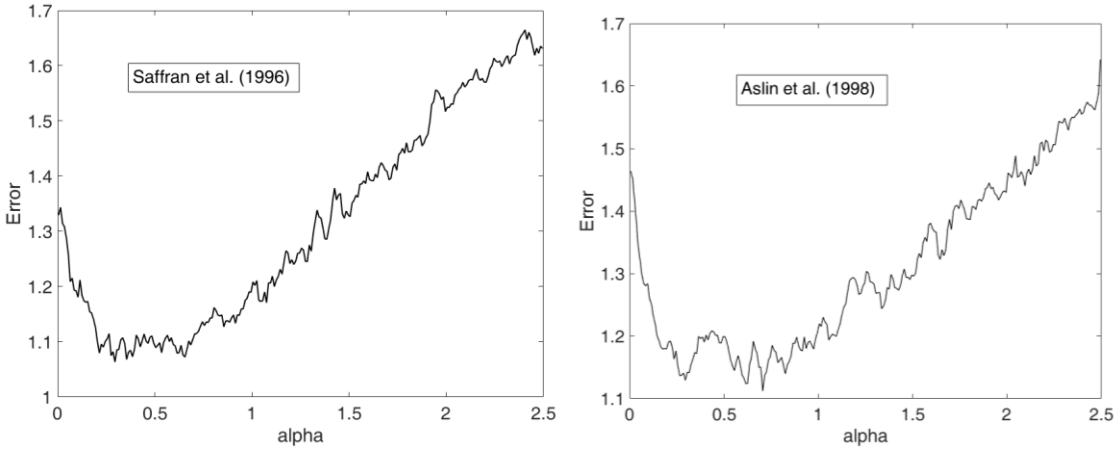


LHS = (1- tanh(αΔ))*Hiddens + (tanh(αΔ))*RHS

(a)


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion better
Saffran et al. (1996)
Aslin et. al (1998)
8 mo. olds
TRACX2
SRN

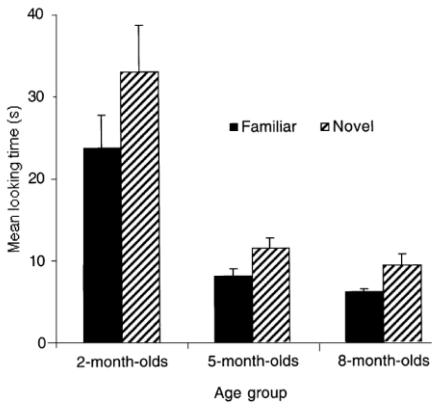
(b)

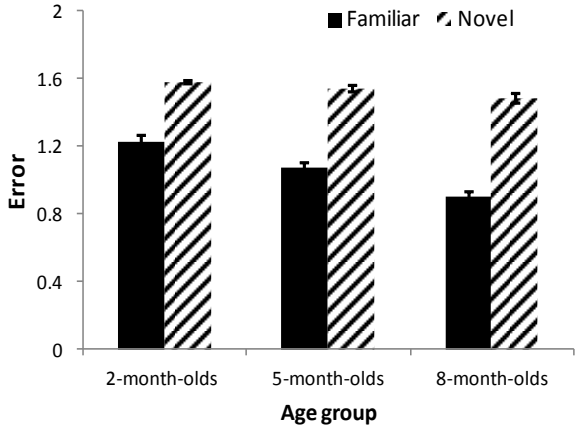

Saffran et al. (1996)
Error
alpha
Aslin et al. (1998)
Error
alpha

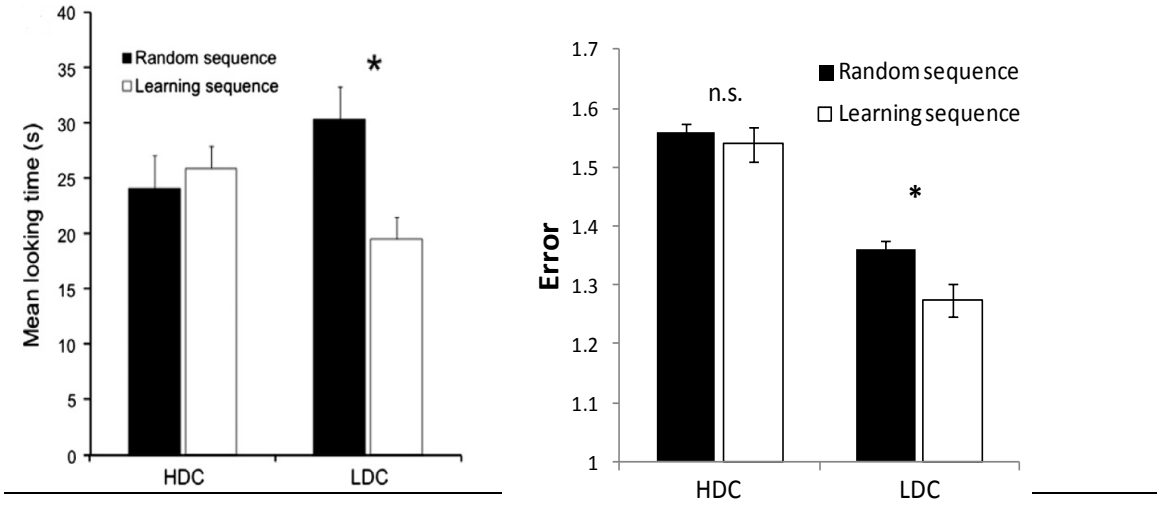
(a)

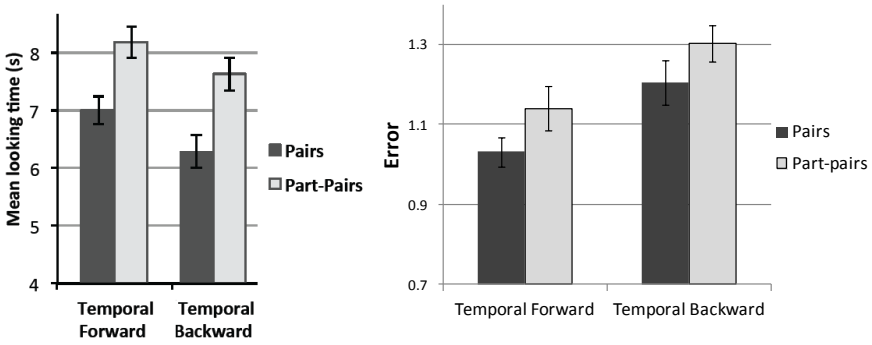
(a) Random Sequence, Each Item is followed by a Different Shape with Equal Probability.

(b) Pair-based Grammar where the Black Bracket Indicates the Fixed Deterministic

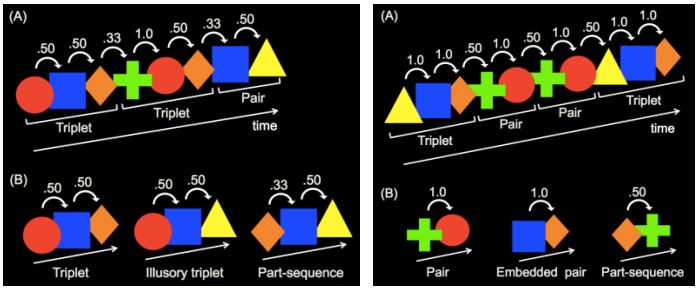




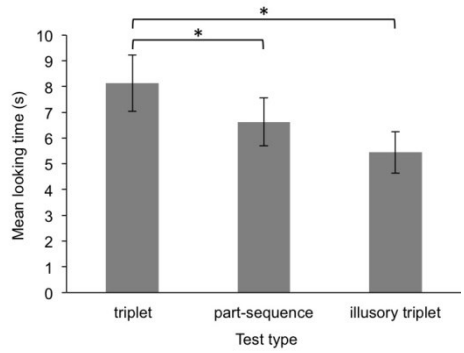

Mean looking time (s)
40
35
30
25
20
15
10
5
0
Random sequence
Learning sequence
*
HDC
LDC
Error
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
n.s.
Random sequence
Learning sequence
*
HDC
LDC

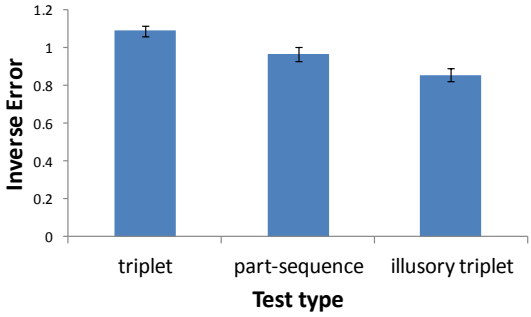

Mean looking time (s)
8
7
6
5
4
Temporal Forward
Temporal Backward
Pairs
Part-Pairs
Error
1.3
1.1
0.9
0.7
Temporal Forward
Temporal Backward
Pairs
Part-pairs

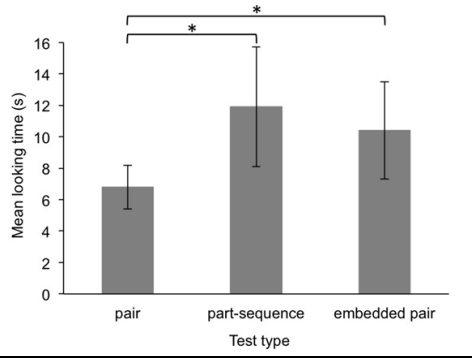
(a)

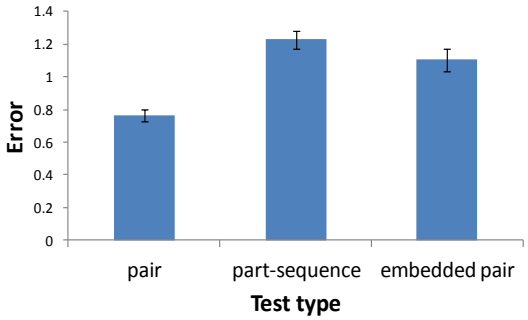

(A)
.50 .50 .33 1.0 .50 .33 .50
Triplet
Triplet
Pair
time
(B)
.50 .50
.50 .50
.33 .50
Triplet
Illusory triplet
Part-sequence
(A)
1.0 1.0 .50 1.0 .50 1.0 .50 1.0 1.0
Triplet
Pair
Pair
Triplet
time
(B)
1.0
1.0
.50
Pair
Embedded pair
Part-sequence

(b)


Mean looking time (s)
0 1 2 3 4 5 6 7 8 9 10
*
*
triplet
part-sequence
illusory triplet
Test type


Inverse Error
0 0.2 0.4 0.6 0.8 1 1.2
triplet
part-sequence
illusory triplet
Test type


Mean looking time (s)
0 2 4 6 8 10 12 14 16
*
*
pair
part-sequence
embedded pair
Test type


Error
0 0.2 0.4 0.6 0.8 1 1.2 1.4
pair
part-sequence
embedded pair
Test type