TRACX2: a RAAM-like autoencoder modeling graded chunking in infant visual-sequence learning

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

[1] LEAD-CNRS UMR 5022, UBFC, Dijon, FR {robert.french@u-bourgogne.fr}
[2]CBCD Birkbeck University of London, UK {d.mareschal@bbk.ac.uk}

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 than all-or-none in nature. As chunks are learned their component parts become more and more tightly bound together. TRACX2 successfully models the data from four experiments from the infant visual statistical-learning literature, including tasks involving 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 over time. One of the greatest challenges for cognition is to find structure within this stream of experiences. Even newborn infants are able to do this (Teinonen, et al. 2009; Bulf, Johnson & Valenza, 2011), and yet, how this is achieved remains largely a mystery.

Two possibilities have been suggested (see Theissen, et al., 2013 for a detailed discussion). The first, characterised as statistical learning, involves using frequency and transition probabilities to construct an internal representation of the regularity boundaries among elements encountered. The second possibility suggests that elements that co-occur are recalled and simply grouped together – or chunked – into single units. 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. What distinguishes these accounts is that the former argues that it is the probabilistic structure of the input sequence that is represented and stored, 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.

We will argue that this is a false dichotomy: both transitional probability learning (statistical learning) and chunking co-exist in one system that smoothly transitions between these apparent 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 are most frequently thought of as all-or-nothing items. Who thinks of "cups" and "boards" when they see the word "cupboard"? Or "foot" and "ball" when they encounter the word "football"? Indeed, chunks like these have essentially the same status as "primitive" words like "boat" or "tree", which are not made of component sub-words. But new chunks do not suddenly appear *ex nihilo* in language. Rather, they are generally formed gradually, their component words becoming more and more bound together with time and usage. For example, 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 of chunks is at the heart of the chunking mechanism in TRACX2.

These ideas were implicit in our initial presentation of the TRACX model (French et al., 2011). In TRACX we showed that a connectionist autoencoder, augmented with conditional recurrence, could extract chunks from a stream of sequentially presented symbols. TRACX

[1] This article is an abridged, modified version of Mareschal, D. & French, R. M. (2017) TRACX2: a connectionist autoencoder using graded chunks to model infant visual statistical learning. *Phil. Trans. R. Soc. B* 2017 372 20160057; DOI: 10.1098/rstb.2016.0057.

had two banks of input units, which it learned 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 learned to form chunks of elements that it recognised as co-occurring (see French et al., 2011 for full details). TRACX successfully captured a broad range of data from the adult and infant auditory statistical learning literature and outperformed existing models of both chunking, notably, PARSER (Perruchet & Vinter, 1998) and statistical learning (SRNs, Cleeremans & McClelland, 1991).

TRACX2 (French & Cottrell, 2014), which we use in this paper to segment and chunk sequential visual items, is an updated version of TRACX. TRACX2 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 conditional jump (i.e. an *if-then-else*) statement from the model, jump statements of this kind not being 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. TRACX (French et al., 2011) and TRACX2 (French & Cottrell, 2014) were used to successfully model the segmentation of syllable (i.e., auditory) streams. In the present article, we use TRACX2 to model four experiments from the infant *visual* statistical learning literature. Visual statistical learning paradigms involve showing infants sequences of looming colored 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 (Kirkham, Slemner & Johnson, 2002) and has yet to be captured by any modeling paradigm.

The TRACX2 Architecture

TRACX2 was initially introduced by French and Cottrell (2014). 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, as shown in the following equation:

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

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*, RHS_t is the activation across the right-hand bank of input nodes, and α is the sigmoid-"steepness" parameter, always set to 1 in the simulations presented here. If at time t, Δ_t is small, this means that the network has learned 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. At each time step, the weights are updated to minimise output error (Fig. 1).

In layman's terms, this means that as you experience

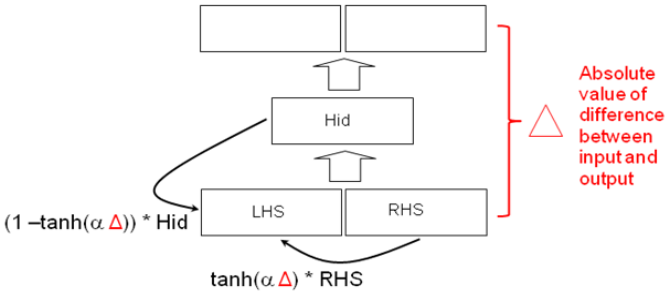


Figure 1. Architecture and information flow in TRACX2. In all simulations reported in this paper, $\alpha = 1$. When Δ is large (items not recognized as having been seen together before on input), almost all contribution to LHS comes from RHS. When Δ is small (items recognized as having been seen together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

items (visual, auditory, tactile) together over and over again, these items become bound to each other more and more strongly into a chunk until we no longer perceive its component parts.

Modeling infant statistical learning

In this section we report on a total of four different simulations using TRACX2 of infant visual statistical learning behaviour. 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. This is a typical parameter used to model age related differences in early learning (e.g., Thomas & Johnson, 2006). 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., Bornstein et al., 1988; Colombo & Mitchell, 2009; Westermann & Mareschal, 2013). Finally, as has been used repeatedly elsewhere, we take network output error as a proxy for looking time in the infant (Mareschal & French 2000; Mareschal, French, Quinn, 2000; Mareschal, Quinn, & French, 2002; Mareschal & Johnson, 2002; French, Mareschal, Mermillod & Quinn, 2004; Westermann & Mareschal, 2013). 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.

We begin by modeling the seminal Kirkham et al. (2002) 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 (Bulf et al., 2011).

Finally, we show that, like 8-month-olds (Slone & Johnson, 2015), TRACX2 forms illusory conjunctions, normally taken as evidence of a statistical learning mechanism, but also shows decreased salience of embedded chunk items, normally taken as evidence of chunking. It, therefore, reconciles two apparently paradoxical behaviours within a single common mechanism.

Visual statistical learning

Kirkham et al (2002) developed a visual analogue of the auditory statistical learning tasks initially developed by Saffran et al. (1996) and Aslin et al. (1998). Instead of listening to unbroken streams of sounds, infants were shown continuous streams of looming colorful shapes in which successive visual elements within a "visual word" were deterministic, but transitions between words were probabilistic (see Fig. 2, leftmost panel). Infants at three different ages were first familiarized 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. found that infants from 2 months of age subsequently looked longer at the random sequence than the structured sequence (even though elements are identical between streams) suggesting that the infants had learned the statistical structure of the training sequence.

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. 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 minimize effects due to input similarity. 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

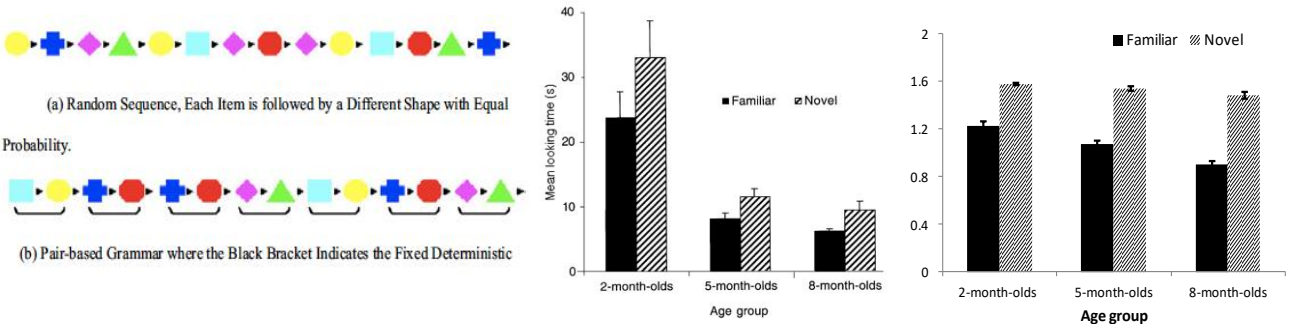


Figure 2. (leftmost panel) Illustration of visual sequences used to test infants (after Addyman & Mareschal, 2013). (middle and rightmost panels) Left-hand panel: Infant performance reported in Kirkham et al. (2002) and, right-hand panel: TRACX2 performance with the familiar structured and novel non-structured sequences. (Error

is the maximum error of the network over all output units; SEM error bars.)

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., 1996), 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 of all ages, looked longer at the randomised sequence than the structured sequence (Fig. 2, rightmost panel).

Visual statistical learning in newborns

Bulf, Johnson, and Villenza (2011) asked whether the sequence-learning abilities demonstrated by Kirkham et al (2002) were present from birth. They tested newborns (within 1 week 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. 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.

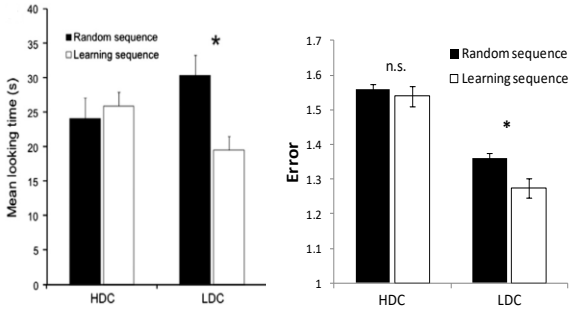


Figure 3. Newborn performance as reported in Bulf & Johnson (2011) in left panel and TRACX2 performance in right panel for familiar structured and novel non-structured sequence.

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 3 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.

Learning embedded and illusory items.

One consequence of chunking is that elements within a chunk become less salient as the chunks are increasingly consolidated. In contrast, statistical learning mechanisms predict that learners should form illusory items from elements that accidentally appear together on occasion. Slone & Johnson (2015) explored 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 underpins infant learning. To do this, they presented 8-months-olds with sequences structured as depicted in Figure 4a. 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 one result suggests that infants utilize chunking, whereas the other results suggests that they engage in statistical learning. TRACX2 captures both of these results well, with the caveat that the model is designed to produce the smallest error on the best learned patterns. (Figs. 4b, 4c). If we consider output error to be a measure of attention (the higher the error, the more attention the infant pays to that item), then we can say that TRACX2 is designed to orient to novel test patterns most (i.e., shows a novelty preference). In short, when modeling a novelty preference, the greater TRACX2's Error on output, the longer the looking time for infants.

Familiarity preferences are, in some sense, the inverse of novelty preferences. This means that the *smaller* the error for an item, the *more* attention the infant pays to that item. Thus, to model familiarity preferences we subtract the error on output from the maximum possible

error and call this "Inverse Error" (Fig. 4c). So, when modeling a familiarity preference, the greater TRACX's Inverse Error, the longer the infants' looking time.

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 (Roder, Bushnell, & Sassville, 2000; Hunter & Ames, 1988; see Sirois & Mareschal, 2004, for a process account of the familiarity-to-novelty shift in a neural network model of habituation). Thus, 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

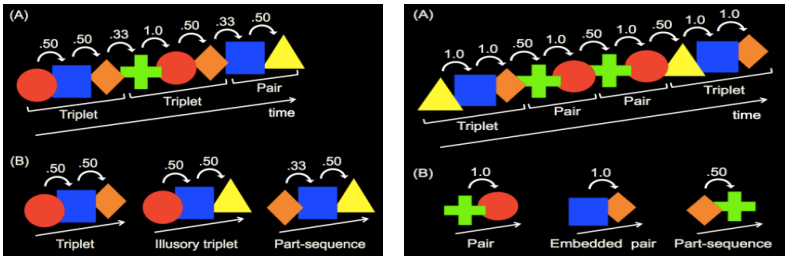
TRACX2 (French & Cottrell, 2014) is an updated version the TRACX architecture (French et al. 2011). As in the original architecture, TRACX2 is a memory-based chunk-extraction architecture. Because it is implemented as a recurrent connectionist autoencoder in the RAAM family of architectures (Pollack, 1989), 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.

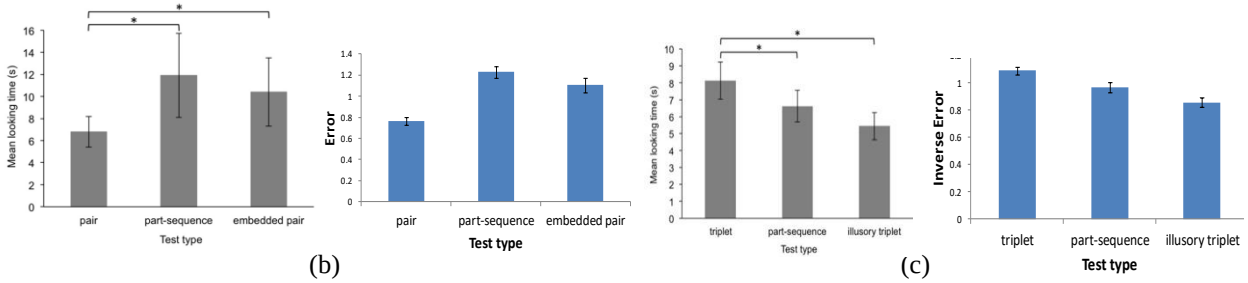
TRACX2 was used to model a representative range of infant visual statistical learning phenomena. No previous models of these behaviours exist. As with the auditory learning behaviours, TRACX2 captures infants' apparent use of forward and backward transitional probabilities, the diminishing sensitivity 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 recognizing previously seen chunks of information. This is a fundamentally different account of infant behaviours than has previously been proposed (Krogh, Vlach & Johnson, 2013).

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; Marcovitch & Lewkowicz, 2009). This would suggest that categorizing learning either as statistical or memory-based is a false dichotomy. Both can happen in a single system, with different behaviours seeming to 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 (Addyman & Mareschal, 2013).

TRACX2 also suggests that there are no specialised mechanisms in the brain dedicated to sequence learning. Instead, sequences emerge from the application of fairly ubiquitous associative mechanisms, coupled with graded top-down re-entrant processing.

Figure 4 (a)





(b) (c)

Figure 4. (a) Familiarisation and testing items for embedded pairs (left panel) and illusory items (right panel) (after Slone & Johnson, submitted). (b) Infant data (left-hand side of figure, familiarity preference, Experiment 1) and TRACX2 performance (right-hand side, SEM error bars). (c) Infant data (left-hand side of figure, novelty preference, Experiment 2) and TRACX2 performance (right-hand panel, SEM error bars). (Figure (a) permission pending).

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.

In conclusion, we believe that chunking cannot be viewed as an all-or-nothing phenomenon. Chunks are learned and over the course of being learned 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 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.

Acknowledgments

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

References

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

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.

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

Bulf, H., Johnson, S. P., & Valenza, E. (2011) Visual statistical learning in the newborn infant. *Cognition, 121*, 127-132.

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

Colombo, J. & Mitchell, D. W. (2009) Infant visual habituation. *Neurobiology of Learning & Memory, 92*, 225–234.

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*.

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.), *Proc of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2016-2221.

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. *JEP:G, 133*, 382-397.

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.

Kirkham, N., Slemmer, J.A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. *Cognition, 83*, B35-B42.

Krogh, L., Vlach, H. A & Johnson, S. P (2013) Statistical learning across development: flexible yet constrained. *Frontiers in Psychology, 3*, art. 598.

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.

Mareschal, D. & Johnson, S. P. (2002) Learning to perceive object unity: A connectionist account. *Developmental Science, 5*, 151-172.

Mareschal, D., French, R. M. & Quinn, P. (2000) A connectionist account of asymmetric category learning in infancy. *Developmental Psychology, 36*, 635-645.

Mareschal, D. & French, R. M. (2000) Mechanisms of categorisation in infancy. *Infancy, 1*, 59-76.

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

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *J. of Mem. and Lang., 39*, 246-263.

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.

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.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996) Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Sirois, S. & Mareschal, D. (2004). An interacting systems model of infant habituation. *J. Cog. Neuro, 16*, 1352-62.

Slone, L. K. & Johnson, S. P. (2015) Statistical and chunking processes in infants' and adults' visual statistical learning. Poster presented and the *Biannual Conf. of the SRCD*, April 2015, Philadelphia, USA.

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

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

Thomas, M. S. C. & Johnson, M. H. (2006). The computational modelling of sensitive periods. *Developmental Psychobiology*, 48, 337-344.

Westermann, G. & Mareschal, D. (2013) From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B, 369*: 201220391