Psychological Review
2011, Vol. 118, No. 4, 614–636

0033-295X/11/$12.00 DOI: 10.1037/a0025255

TRACX: A Recognition-Based Connectionist Framework for Sequence Segmentation and Chunk Extraction

Robert M. French
LEAD-CNRS/Université de Bourgogne

Caspar Addyman
LEAD-CNRS/Université de Bourgogne and
Birkbeck University of London

Denis Mareschal
Birkbeck University of London

Individuals of all ages extract structure from the sequences of patterns they encounter in their environment, an ability that is at the very heart of cognition. Exactly what underlies this ability has been the subject of much debate over the years. A novel mechanism, implicit chunk recognition (ICR), is proposed for sequence segmentation and chunk extraction. The mechanism relies on the recognition of previously encountered subsequences (chunks) in the input rather than on the prediction of upcoming items in the input sequence. A connectionist autoassociator model of ICR, truncated recursive autoassociative chunk extractor (TRACX), is presented in which chunks are extracted by means of truncated recursion. The performance and robustness of the model is demonstrated in a series of 9 simulations of empirical data, covering a wide range of phenomena from the infant statistical learning and adult implicit learning literatures, as well as 2 simulations demonstrating the model's ability to generalize to new input and to develop internal representations whose structure reflects that of the items in the input sequence. TRACX outperforms PARSER (Perruchet & Vintner, 1998) and the simple recurrent network (SRN, Cleeremans & McClelland, 1991) in matching human sequence segmentation on existing data. A new study is presented exploring 8-month-olds' use of backward transitional probabilities to segment auditory sequences.

Keywords: chunk extraction, statistical learning, implicit learning, recursive autoassociative memory, autoassociators

Supplemental materials: http://dx.doi.org/10.1037/a0025255.supp

Individuals of all ages extract structure from the sequences of patterns they encounter in their environment, an ability that is at the very heart of cognition. Exactly what underlies this ability has been the subject of much debate over the years. One of the most widely accepted proposed explanatory mechanisms is learning based on prediction. The idea is that individuals are constantly engaged in predicting upcoming patterns in their environment based on previously encountered patterns. Learning, in this view, is a process of gradually aligning these predictions with the outcomes that actually occur. Prediction-driven learning is the cornerstone of numerous computational models of sequence processing and, in particular, underlies the very well known simple recurrent network (SRN, Elman, 1990). In this article, we propose an alternative mechanism, *implicit chunk recognition* (ICR), a process that is based not on prediction but rather on the recognition of previously (and frequently) encountered subsequences of patterns (chunks).

This does not mean that prediction plays no role in sequence processing or in cognition more generally. But it turns out that prediction-driven models, in general, and the SRN, in particular, cannot account for a number of recent results in infant statistical learning (SL) and adult implicit learning (IL). This strongly suggests that there must be some other processes underlying sequence processing and chunk extraction. We have developed a connectionist implementation of ICR, the truncated recursive autoassociative chunk extractor model (TRACX), which is able to handle empirical data that are problematic for prediction-based models. In addition, we show that TRACX accounts for a wide range of other results in sequence segmentation and chunk extraction.

Robert M. French, Centre National de la Recherche Scientifique, Laboratoire d'Etude de l'Apprentissage et du Développement, Unité Mixte de Recherche 5022 (LEAD-CNRS UMR 5022), Département de Psychologie, Université de Bourgogne, Dijon, France; Caspar Addyman, LEAD-CNRS UMR 5022, Département de Psychologie, Université de Bourgogne, and Centre for Brain and Cognitive Development (CBCD), Department of Psychological Science, Birkbeck University of London, London, England; Denis Mareschal, CBCD, Department of Psychological Science, Birkbeck University of London.

This work was made possible in part by European Commission Grant FP6-NEST-029088, French Agence Nationale de la Recherche Grant ANR-10-065-GETPIMA, and United Kingdom Economic and Social Research Council Grant RES-062-23-0819 under the auspices of the Open Research Area France–United Kingdom funding initiative. We thank Pierre Perruchet for his many insightful comments on the work in this article, his assistance in the preparation of the familiarization sequence and test stimuli for the replication of Aslin et al. (1998), and his allowing us to use his input-data encodings for a number of the simulations presented in this article.

Correspondence concerning this article should be addressed to Robert M. French, LEAD-CNRS UMR 5022, Pôle AAFE, Université de Bourgogne, Dijon, France. E-mail: robert.french@u-bourgogne.fr

Models of sequence processing and chunk extraction fall, broadly speaking, into three categories: symbolic/hybrid models, connectionist models, and normative statistical models. PARSER (Perruchet & Vintner, 1998, 2002) is the best known model from the first class, and the prediction-driven SRN (Cleeremans, 1993; Cleeremans & McClelland, 1991; Elman, 1990; D. Servan-Schreiber, Cleeremans, & McClelland, 1991) is the best known from the second class. We examine both of these models in some detail and compare their performance and architectures with that of TRACX. We also include a succinct comparison of TRACX with a number of Bayesian models discussed in Frank, Goldwater, Griffiths, and Tenenbaum (2010).

In the remainder of this article, we illustrate the TRACX framework by focusing on adult IL and infant SL, two domains that involve sequence segmentation and chunk extraction. Perruchet and Pacton (2006) suggested that although the former area tends to emphasize *recognition* of remembered wordlike chunks and the latter emphasizes *prediction* based on statistical distribution information, these two historically distinct fields of enquiry may in fact constitute two approaches to the same problem. They went on to suggest that providing a unified framework for these two areas is the major theoretical challenge facing researchers in both fields. We suggest that the TRACX model, a recursive connectionist autoassociator, not only provides a simple, parsimonious, unifying, memory-based solution to the challenge posed by Perruchet and Pacton (2006) but also provides a general, recognition-based connectionist framework for chunk extraction from any sequence of patterns.

We first examine the information that underlies chunk extraction in these two domains. People exploit multiple cues in word segmentation (Christiansen, Allen, & Seidenberg, 1998; Cunillera, Camara, Laine, & Rodriguez-Fornells, 2010; Perruchet & Tillman, 2010); therefore, a computational model of sequence segmentation must be able to do likewise. We then review the existing computational models of sequence segmentation and chunk extraction, notably SRN and PARSER, and suggest that when doing sequence segmentation, these models differ fundamentally in their reliance on prediction versus recognition/chunking. The TRACX model is then introduced as a recognition-based connectionist model of chunk extraction that can use multiple statistical cues to extract and form chunks. We show that TRACX can successfully capture performance on eight different studies of adult and infant sequence segmentation, can scale up to large corpora, and can generalize novel input in meaningful ways. Moreover, it succeeds in capturing human performance when the SRN model and PARSER seem to fail.

Adult IL

Shanks (2005, 2010) listed a number of diverse examples of behaviors that might be considered to demonstrate IL. Tasks and measures that show IL include artificial grammar learning (AGL, Reber, 1967), faster responses in serial reaction time tasks (Destrebecqz & Cleeremans, 2001), classical conditioning of eye blink responses (Clark & Squire, 1998), unconscious position bias in a qualitative judgment task (Nisbett & Wilson, 1977), and dynamic process control tasks (Berry & Broadbent, 1984).

Furthermore, Cleeremans and Dienes (2008) emphasized three key properties of IL: (a) that learning happens without intention or effort, (b) that it happens without the participant's awareness, and (c) that the resulting knowledge is unconscious. IL research has mainly focused on sequence segmentation tasks, in particular, the original AGL task introduced by Reber (1967) and the serial reaction time (SRT) task developed by Nissen and Bullemer (1987). Both experimental paradigms afford the experimenter a high degree of control because the type of stimuli and their relations to each other can be precisely chosen.

Chunking in Adult IL

Perruchet and Pacteau (1990) presented convincing empirical evidence that simple associative learning can drive the formation of chunks and chunk-fragments in AGL. In a series of three experiments, they showed that grammaticality judgments could be based on an explicit knowledge of valid bigrams. They found that participants who were shown only a list of valid bigrams rather than full grammatical sentences containing these bigrams were equally accurate at grammatically judgments. In other words, their grammaticality judgments could have simply been based on the recognition of the bigrams, a fact that was confirmed in the final experiment that directly tested bigram recognition. Perruchet and Pacteau concluded that chunk knowledge is explicit, a position similar to that of Dulany, Carson, and Dewey (1984), who hypothesized that AGL performance was attributable to participants' possessing and consciously applying an incomplete set of micro-rules. The TRACX model relies entirely on the recognition of recurring chunks in the input stream while eschewing the need for explicit awareness.

Finally, a recent study by Giroux and Rey (2009), whose results are simulated in this article, provides evidence that when chunks are learned, the subunits making up these chunks are forgotten unless they are refreshed independently. This would imply that chunks are encoded as atomic entities rather than as associations between their constituent elements. This process, known as lexicalization, means that a chunk, once fully formed, is thereafter treated as an indivisible word, causing any words within it to be overlooked. For example, we fail to notice the words *break* and *fast* in *breakfast* or *cup* and *board* in *cupboard*. Giroux and Rey (2009) demonstrated this effect in an artificial grammar task. Dahan and Brent (1999) and Perlman, Pothos, Edwards, and Tzelgov (2010) have reported similar results.

Evidence for the Use of Transitional Probabilities (TPs) in Adult Sequence Segmentation

The TP between two syllables, *X* and *Y*, is the probability that given one of the syllables, the other will also occur. They come in two types: *forward TPs* and *backward TPs.* A concrete example, with letter combinations in English and French, will serve to illustrate the difference between them. Consider the bigram *qu*. The forward TP between its two letters is the probability that given a *q* in the first position, a *u* will follow. In English, this probability is, for all intents and purposes, 1. The backward TP is the probability that given a *u* in the second position, it will be preceded by a *q*. In English, this backward TP, which is approximately .01, is considerably lower than the forward TP. Backward TPs can, in some cases, be considerably higher than forward TPs. Consider the *ez* suffix in French (as in "parl*ez*-vous français?"). The probability

that given a *z*, it will be preceded by an *e* is .84 (http://www.lexique.org/listes/liste_bigrammes.php), whereas the probability that an *e* will be followed by a *z* is a mere .027. Sensitivity to forward TPs between word syllables is frequently cited as the most important cue used by infants in word segmentation (e.g., Aslin, Saffran, & Newport, 1998; Saffran, 2001; Saffran, Aslin, & Newport, 1996; Thiessen et al., 2005), but Pelucchi, Hay, and Saffran (2009a) have recently shown that infants are also sensitive to backward TPs.

There is also evidence for the use of TPs in adult AGL. Saffran, Newport, Aslin, Tunick, and Barrueco (1997) demonstrated that adults could use forward TPs to segment an AGL. Perruchet and Desaulty (2008) demonstrated that adults use both forward and backward TPs for word segmentation. Their study showing the use of backward TPs as cues for segmentation will be of particular importance in the present article because it provides a comparison of recognition-based TRACX and prediction-based SRNs (Cleeremans & McClelland, 1991; Elman, 1990). We show that standard implementations of an SRN are not able to simulate data in which segmentation cues arise from backward TPs, whereas TRACX has no problem doing so.

Summary of Adult Sequence Segmentation

During the IL of sequential information, adults are sensitive to syllable and chunk frequencies in the input stream, as well as forward and backward TP information. They also gradually extract stimulus chunks from a continuous syllable stream and gradually forget the subunits within chunks if these subunits are not refreshed outside the context of the larger chunk in which they are found.

SL in Infants

Infant SL initially focused largely on the question of how infants segment a continuous speech stream into lexical units. The raw auditory signal generated by human speech is notoriously hard to segment into words because breaks in the continuity of the signal are poorly correlated with actual word boundaries (Cole & Jakimik, 1980) and, therefore, are poor cues for word segmentation. Nonetheless, it has been shown repeatedly that infants as young as 8 months can discriminate between items based on their novelty with respect to a familiarization corpus of training items (e.g., Aslin et al., 1998; Jusczyk & Aslin, 1995; Pelucchi et al., 2009a, Pelucchi, Hay, & Saffran, 2009b; Saffran et al., 1996).

Infants have been shown to be sensitive to many different cues that could arguably contribute to their ability to segment natural language. These include statistical regularities arising from grammar; patterns of rhythm, stress, and tempo; and duration of phonemes and syllables (Cooper & Paccia-Cooper, 1980; E. K. Johnson & Jusczyk, 2001; Jusczyk, Hohne, & Bauman, 1999; Jusczyk, Houston, & Newsome, 1999; Klatt, 1976; Mattys, Jusczyk, Luce, & Morgan, 1999; Nakatani & Shaffer, 1978). However, speech segmentation can still occur, even in the absence of most of these cues (Aslin et al., 1998; Jusczyk & Aslin, 1995; Saffran et al., 1996). In particular, Saffran et al. (1996) focused on syllable co-occurrence probabilities within the sound stream; specifically, the difference between the forward TPs between adjacent syllables within words, versus TPs between adjacent syllables on either side of word boundaries. Saffran et al. (1996) used a synthesized speech stream consisting of a random sequence of words drawn from a small corpus of trisyllabic words that were chosen so as not to resemble any words in the language to which the infant was normally exposed. With very short exposure (2 min), infants were still able to discriminate words from nonwords and partwords.[1] The words and partwords that the infants were tested on were present in the training stream, so the authors reasoned that the discrimination was based on the fact that adjacent-syllable TPs were lower at word boundaries than within words.

In a further study, Aslin et al. (1998) controlled for bigram frequency and found that infants were still able to discriminate words from partwords, reinforcing the conclusion that the crucial cue for segmentation was adjacent-syllable TP information. A large number of other results have supported these initial findings (e.g., Saffran, 2001; Saffran, Johnson, Aslin, & Newport, 1999; Saffran & Wilson, 2003; Thiessen & Saffran, 2003, 2009). Even in the visual domain, infants appear to show sensitivity to co-occurrence statistics with sequential stimuli (Kirkham, Slemmer, & Johnson, 2002) and to conditional probabilities with spatially distributed visual stimuli (Fiser & Aslin, 2001). Intriguingly, Marcovitch and Lewkowicz (2009) reported that TP and frequency information both contribute independently to infant segmentation of visual sequences. Finally, infant SL has also been demonstrated using speech stimuli in a natural language unfamiliar to the infants (Pelucchi et al., 2009a, 2009b).

Jusczyk, Houston, and Newsome (1999) demonstrated that lexicalization of chunks also takes place, even in infants. They showed that 9-month-olds learning the bisyllabic words *doctor* or *candle* do not subsequently respond to the single syllable words *dock* or *can*. A similar result was found in adults by Giroux and Rey (2009).

Summary of Infant SL

The difference in TPs for syllable-pairs within words and TPs for syllable-pairs astride word boundaries has been shown to be a very powerful cue in infant word segmentation. This has been demonstrated both for forward and for backward TPs. This sensitivity to co-occurrence statistics has been shown for simple, carefully controlled artificial languages, streaming visual sequences, and natural language speech streams. Finally, as in adults, chunked words appear to have a relatively atomic nature, and subchunks making up those words, if they are not heard independently elsewhere, will be forgotten.

Computational Models of Sequence Segmentation and Chunk Extraction

We divide the existing computational models of sequence segmentation into three classes: symbolic/hybrid, connectionist, and descriptive statistical models. This classification is designed only for broad organizational purposes, since many of the models share aspects of more than one class.

[1] A partword is a syllable cluster consisting of syllables from the end of one word and the beginning of the following word. So, for example, *yellow basket* would give rise to the partword *lowbas*.

Symbolic/Hybrid Models

One of the most successful approaches to modeling adult AGL has come from symbolic/hybrid models. These include the competitive chunking network (CCN, E. Servan-Schreiber & Anderson, 1990, see also Boucher & Dienes, 2003), PARSER (Perruchet & Vintner, 1998), and CLARION (Sun, 1997), which are built on principles drawn from both connectionist and symbolic modeling and from models based on the ACT-R framework (Lebiere, Wallach, & Taatgen, 1998). Of these, PARSER has modeled the widest range of results, and PARSER is, without question, one of the most successful models of chunk extraction to date. For this reason, its architecture is worth a more detailed presentation.

PARSER consists of a working memory (WM) structure that initially contains the individual syllables of the language. The program serially scans an input stream of syllables, chunking adjacent syllables and storing these chunks in WM. Syllable chunks are formed at each time step of the program by randomly selecting a value of 1, 2, or 3, which will determine the number of successive items of the sequence the program will chunk together on that time step. Items making up a chunk can be either primitives (i.e., individual syllables) or other chunks that are still above a preset strength threshold in WM. If the chunk already exists in WM, then it is strengthened; if the chunk is not already present in WM, then it is placed in WM. Now, suppose that the program decides to pick the next two items in the sequence to form a chunk. If the incoming syllables are, say, *cdabefcdabcd*—, it will scan its WM to find the two longest chunks stored in WM that could make up the initial section of *cdabefcdabc*—. If these are, say, *cd* and *abef*, it will create a new chunk, *cdabef,* and put it in WM. The new chunk interferes with all other chunks in WM based on how much syllable overlap it shares with them, giving this particular aspect of PARSER, perhaps inadvertently, a somewhat connectionist feel. A forgetting parameter is also applied to all chunks in WM at every time step. Only chunks that are above a particular predetermined strength threshold are allowed to remain in WM. This ensures that PARSER's WM never contains too many chunks.

Connectionist Models

Connectionist modelers have typically taken one of two basic approaches to sequential learning, using either recurrent or nonrecurrent networks. These latter models of AGL have been somewhat less successful, so there have only been a few attempts to apply them. For example, Dienes (1992) used an autoassociator with localist inputs and no hidden layer that attempted to reproduce entire sentences. For modeling infant SL, Aslin, Woodward, LaMendola, and Bever (1996) used a three-layer feedforward network with a three-item moving window of inputs that attempted to predict the presence of boundaries. However, this model only worked when phonemes were encoded by their features and with short (3–5 word) sentences. Sirois, Buckingham, and Shultz (2000) used a simple autoassociator model to show how structural regularities extracted from a speech stream by infants could be transferred to a different set of sounds with different surface features (see Marcus et al., 1999). Finally, B. Anderson (1999) produced a Kohonen-network model of Saffran et al. (1996). Presumably, these models were not developed further because they only had limited success on the problems to which they were applied.

However, there is an extensive body of research with SRNs (Elman, 1990) for sequence processing. SRNs are prediction-driven feedforward–backpropagation networks that attempt at each time step to predict the next item of the sequence based on the current item on input plus the network's own internal state on the previous time step. Elman's original article, however, did not propose SRNs as a word-learning mechanism. Cleeremans and McClelland (1991) were the first to specifically apply the SRN to capture sequence-learning performance. Other authors (e.g., Altmann, 2002; Cairns, Shillcock, Chater, & Levy, 1994, 1997; Christiansen, 1999; Christiansen, Allen, & Seidenberg, 1998; Mirman, Graf Estes, & Magnuson, 2010) later attempted to use SRNs to segment words from continuous speech streams of infant-directed speech. Dominey and colleagues (e.g., Dominey, 1998; Dominey & Ramus, 2000) proposed replacing SRNs with more biologically realistic temporal recurrent networks (TRNs) in which each time-step corresponded to 5 ms of real time and the hidden and recurrent units were leaky integrators with response times between 20 ms and 400 ms. This family of models was successfully applied to serial-reaction time data (Dominey, 1998), the Saffran et al. (1996) data that showed that 8-month-old infants learned words better than partwords and could also discriminate the distinctive rhythmic consonant-vowel patterns similar to those found in English (Nazzi, Bertoncini, & Mehler, 1998) The underlying principle of both SRNs and TRNs is the same—namely, the prediction of future events—as opposed to memory-based models that simply recall whether a particular item (or sequential cluster of items) has been encountered before.

Normative Statistical Models

Whereas the previous models (e.g., PARSER, SRN) attempt to model the *processes* underlying word segmentation, normative statistical models, in general, do not. Instead, they generally view segmentation as a problem of descriptive statistics. For example, an analysis of the infant-directed speech in the CHILDES database (MacWhinney & Snow, 1985) by Christiansen, Onnis, and Hockema (2009) revealed that there is ample TP and word boundary distribution information in the naturally occurring speech that infants hear to support lexical segmentation. Precursors to this approach are found in the early work of Harris (1954, 1955), who found that patterns of phoneme distributions in English gave information that could, in principle, be used to predict word boundaries. The most recent and sophisticated version of this approach has come from Swingley (2005). As with Harris's (1954, 1955) original work, this is a formal statistical analysis rather than a process model of human cognitive performance. Swingley (2005) described the success of a particular statistical algorithm applied to infant-directed speech corpora in English (Korman, 1984) and Dutch (van de Weijer, 1998). This algorithm is syllable-based and uses a heuristic clustering algorithm based on n-gram frequencies and pointwise mutual information. Swingley compared this algorithm with an approach based on n-gram frequencies alone, without accounting for conditional probability, and found that combining both types of cues resulted in a very large benefit for both accuracy and comprehensiveness of segmentation performance. The conclusion of this descriptive statistical analysis strongly suggests that any process model of word segmentation should also incorporate both types of cues.

Another normative approach based on descriptive statistics comes from a Bayesian perspective. There are numerous approaches of this type (e.g., de Marcken, 1995; Brent & Cartwright, 1996; Brent, 1999; Goldwater, Griffiths, & Johnson, 2009; Robinet & Lemaire, 2009; Venkataraman, 2001), most recently, Frank et al. (2010). Originally, these approaches were described in terms of minimum description length, but more recently, they have been framed in terms of maximum likelihood estimation. On a conceptual level, all approaches of this type operate in a similar fashion (e.g., Brent, 1999). They are not process models and use batch processing of the data. That is, they take the whole corpus as a single, unsegmented stream of phonemes and attempt to redescribe it in terms of a set of words, thereby forming a more compact representation of the stream. We discuss the performance of this class of models, as reported by Frank et al. (2010), in more detail below.

Summary of Computational Models of Sequence Segmentation and Chunk Extraction

Three major computational approaches—symbolic/hybrid, connectionist and descriptive statistical—have attempted to account for human sequence processing performance, in particular, in the areas of SL and IL. One of the key features that separates symbolic/hybrid models (e.g., PARSER and CCN) from connectionist models (e.g., SRN and TRN) is that the former rely on the *recognition* of already encountered items, whereas are latter are based on the *prediction* of upcoming items in the input stream. While normative statistical approaches, such as Swingley's (2005) algorithm, can provide good accounts of the putative high-level computations required for word segmentation, unlike TRACX, PARSER, or SRNs, they are not process models and are, therefore, difficult to compare directly with process models. For this reason, we have not concentrated on these models in the present article (see Luce, 1995; Mareschal & Westermann, 2010, and McClelland et al., 2010, for further discussion of process versus normative models).

The TRACX Model

We propose a simple recognition-based connectionist model of sequence segmentation, TRACX, which provides a unifying framework not only for infant speech segmentation and adult IL but also for sequence segmentation and chunk extraction, more generally. It combines the features of chunk-based models into a single, parsimonious connectionist architecture.

The TRACX Architecture

TRACX has its roots in the recursive autoassociative memory (RAAM) family of architectures (Blank, Meeden, & Marshall, 1992; Pollack, 1989, 1990). It is a connectionist autoassociator model based on the implicit chunk recognition of previously encountered subsequences of acoustic primitives, in this case, syllables or phonemes. We have chosen the expression "implicit chunk recognition" because, unlike various other connectionist models of chunk extraction, such as SRNs (Cleeremans & McClelland, 1991) and TRNs (Dominey & Ramus, 2000), TRACX is based on the recognition of previously encountered subsequences of items rather than on the prediction of upcoming items. This recognition is implicit because, unlike symbolic/hybrid models of chunk extraction, such as PARSER (Perruchet & Vintner, 1998) and CCN (Boucher & Dienes, 2003), there is no explicit storage of symbolic information in a separate WM. As in all connectionist models, information is stored in the synaptic weights of the network and, as such, is not directly accessible to the system.

Autoassociators are neural networks that gradually learn to produce output that is identical to their input. Items that they have encountered frequently will be reproduced on output; items that have never been encountered before or that have been encountered only infrequently will produce output that does not resemble the input. In other words, an autoassociator provides a simple way of answering the question, "Has the current input been encountered frequently before?" If the error on output is small (i.e., the network has already successfully learned the input pattern) then the network "concludes" that it has encountered that input pattern before. If the error on output is large, this means that the network does not remember having previously encountered the current input. This fact is essential to the TRACX approach.

Autoassociators have been used in the computational modeling of behavior at least since Anderson's "brain-state-in-a-box" model (BSB, J. A. Anderson, Silverstein, Ritz, & Jones, 1977). Their psychological and biological plausibility is now well established (Rolls & Treves, 1997), and they have been successfully used as psychobiologically plausible models of face perception (Cottrell & Metcalfe, 1991; Dailey, Cottrell, Padgett, & Adolphs, 2002), hippocampal/episodic memory (Gluck & Granger, 1993; Gluck & Meyers, 1997), serial recall memory (Farrell & Lewandowsky, 2002), infant categorization (French, Mareschal, Mermillod, & Quinn, 2004; French, Mermillod, Quinn, & Mareschal, 2001; Mareschal & French, 2000; Mareschal, French, & Quinn, 2000), and infant habituation (Sirois & Mareschal, 2004). Finally, autoassociative networks can be shown to scale, from the restricted experimental paradigms used in laboratory tasks to very large data sets.

The original autoassociators (e.g., J. A. Anderson et al.'s, 1977, BSB) had no hidden layer and, therefore, developed no internal representations of the data that were input to them. A hidden layer was added later in order to use autoassociators as data-compression networks (e.g., Cottrell, Munro, & Zipser, 1988). Mareschal and French (2000), Mareschal, French, and Quinn (2000), and French et al. (2004, 2001) used autoassociators with a hidden layer to model infant categorization. TRACX is an autoassociator with a hidden layer that actively uses the compressed representations developed by its hidden layer. In the case of word segmentation, we suggest that infants, as well as adults, are involved in a continual process of encoding new input into internal representations and assessing this input against those representations, thereby allowing them to discriminate between words and partwords in the input stream. For infants and adults, we assume that their behavior will be different for syllable sequences that they recognize as having heard before, compared with those that they do not recognize as having heard before (e.g., Aslin et al., 1998; Giroux & Rey, 2009; Jusczyk & Aslin, 1995; Pelucchi et al., 2009a, 2009b; Perruchet & Desaulty, 2008; Saffran et al., 1996). Similarly, TRACX will produce a smaller error on output for items that it recognizes as having been encountered before (i.e., are

better learned), compared with items that it recognizes less well (i.e., are less well learned).

TRACX uses only standard learning parameters from the connectionist literature, combined with a mechanism for presenting internal representations on input. It performs chunking by "asking itself" whether it has previously encountered the two items currently on its input together. If it has, then it chunks them. As we show below, TRACX was able to successfully replicate the results of seven different published studies from the adult IL and infant SL literature, in addition to predicting the results of our own study on equal TPs with different word/partword frequencies. In addition, we demonstrate that it can scale up to large databases, in this case, of infant-directed language. And finally, we show that in a simple bilingual microlanguage acquisition task, it is capable of generalizing to new data and of developing clusters of internal representations that reflect the structure of the data it is processing.

TRACX Implementation

The implementation and operation of TRACX is extremely simple. It is a three-layer feedforward–backpropagation[2] autoassociator (see Figure 1). A syllable stream, *S*, is input syllable by syllable to the network. Each of *N* syllables is locally coded as a vector of *N* bipolar values (i.e., the bipolar representation for the *k*th syllable has a 1 in the *k*th position and has −1 elsewhere). TRACX has a 2*N*–*N*–2*N* feedforward-only topology. Importantly, the number of hidden units must be equal to half of the overall length of the input layer in order to allow recursion. The input nodes are organized into a left-hand (LH) side and a right-hand (RH) side, each side designed to hold vectors of length *N*. *N* is determined by the number of syllables used in each task and, in our simulations, can vary from 12 to 90.

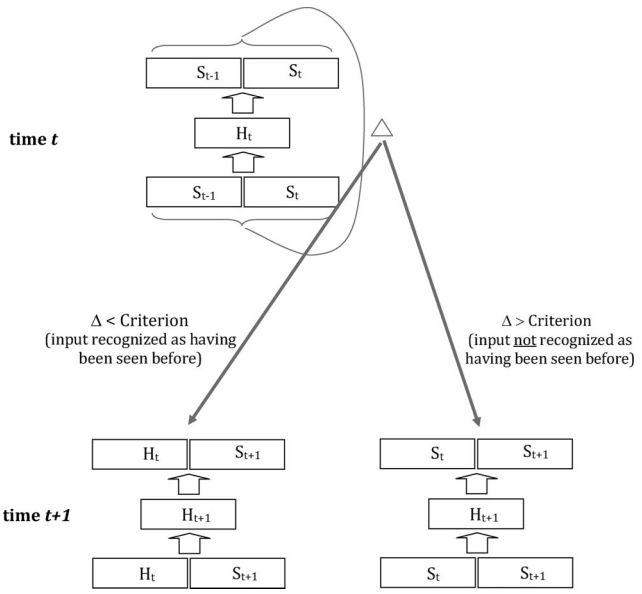


Figure 1. The TRACX model is a 2N–N–2N feedforward–backpropagation autoassociator. If Δ, the difference between the input activations and the output activations, is below criterion at time *t*, the hidden unit vector will become the next left hand (LH) vector of the input at time *t* + *1* and the right hand (RH) vector will be next item in the list. If, on the other hand, the input–output activation Δ at time *t* is large, the LH vector of the input at time *t* + *1* will be the input vector at time *t*, and the RH input vector will be the next item in the list. H_t = hidden-layer activation vector at time *t*; S_t = bipolar representation of syllable in the input stream, S, that is presented to the network at time *t*.

Both input and hidden layers include a bias node whose activation is permanently set to −1. The learning rate of the network is set to .04 for eight of the 10 simulations. In two other simulations, lower learning rates were used for reasons that are explained in the discussion of these simulations. The momentum term was always set to 0. A Fahlman offset of .1 was added (Fahlman, 1988) to the standard backpropagation algorithm to prevent learning stagnation. This is a constant value added to the slope of the error landscape to speed learning and is in no way fundamental to the architecture. A standard hyperbolic tangent function (i.e., $h(x) = \frac{1 - e^{-\beta x}}{1 + e^{-\beta x}}$, with $\beta = 1$) is used on the hidden and output layers. Syllable representations from the input stream are fed into the input layer sequentially. The network error at time *t* is determined by comparing the output of the network at time *t* with its input at the same time *t*. The error is considered to be the maximum error over all of the output nodes, a measure particularly suited to localist representations.

We designate the LH input to the network at time *t* as LH(*t*), the input to the RH side of the network at time *t* as RH(*t*), and the hidden-layer activation vector at time *t* as H(*t*). The syllable in the input stream that is presented to the network at time *t* is designated *S*(*t*) (See Figure 1). Note that the left–right distinction does not reflect any spatial ordering of the incoming information but simply reflects its *temporal* ordering. On each time step, RH(*t*) is *S*(*t*). The error criterion chosen was .4. Thus, for an output to be considered below criterion, the error on each output node must be below .4.[3]

Information flows through the network as follows:

1. Initially, there are two syllables on input.
2. Activation is propagated forward, resulting in activation on the output nodes.
3. The output activation is compared with the current input, and an error measure is thereby determined.
4. This error measure (the maximum error over all output nodes) is compared with the error criterion.
5. If, on the one hand, the error is greater than the error criterion, then the next input to the network is created by moving the value of RH(*t*) to the LH side of the input, LH(*t* + 1). The next element in the input stream, S(*t* + 1) is put into the RH side of the input RH(*t* + 1).

[2] Although backpropagation has sometimes been criticized as not being a neurobiologically plausible learning mechanism (Crick, 1989), O'Reilly and Munakata (2000) have presented a detailed defense of this class of models, showing that they are in practice isomorphic to contrastive Hebbian learning models, which can be justified in a neurobiologically plausible manner.

[3] An output error less than .4 for all units of a bipolar encoding means that the network can round all values to their correct value.

6. If, on the other hand, the output error is below the error criterion, then the hidden-unit values, $H(t)$, are put into the LH side of the input, LH($t + 1$). The next element in the input stream, $S(t + 1)$, is put into the RH side of the input, RH($t + 1$).

7. Activation is propagated forward again, and processing continues as from 3, above.

Learning in TRACX

The RH input units always contain an element from the sequence (i.e., an item from the environment). However, the LH input units can contain either an item from the sequence or the network's internal representation of a number of items. These are very different entities, one coming to the input directly from the sensory interface, the other being internally generated and coming to the input from within the network.

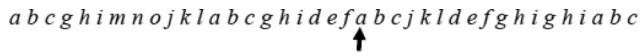
There is now ample evidence that signals within the brain can be either internally generated or externally generated through direct perception. In general, people are good at differentiating these two classes of information (e.g., Christoff, Ream, Geddes, & Gabrieli, 2003; M. K. Johnson & Raye, 1981), suggesting that the signals can be differentiated internally. Moreover, internally generated signals can lead to actual learning, especially in the context of memory consolidation and reactivation (e.g., Gelbard-Sagiv, Mukamel, Harel, Malach, & Fried, 2008) and motor control (e.g., Jeannerod, 1995). However, the learning from internally generated signals is often less potent than from externally generated signals (e.g., Holmes & Calmels, 2008), possibly to avoid overlearning previously encountered, low-frequency items or memories or to avoid overlearning nonadaptive behaviors. Consequently, we made the additional assumption that learning only occurred on 25% of the events in which there is an internally generated representation (chunk) in the LH units than when there is a real item from the environment in the LH units. Another way of thinking about this is that the system pays greater attention (in terms of learning) to new information arising from outside the system than to internally generated information arising from the recognition of previously encountered items.

Testing the Performance of TRACX

Testing consists of presenting each item to be tested to the network. Consider first a two-syllable word, *ab*. Encodings of *a* and *b* are put into the network's LH and RH inputs and activation propagated through the network. The error for *ab* is the maximum error over all output units. Now consider a three-syllable item, *abc*. As before, the first two syllables, *a* and *b*, are put into the network's LH and RH input units, respectively, and the resulting activation is fed through to the network's hidden layer. The hidden-layer representation produced by *a* and *b* on input is then put into the LH input units, and *c* is put into the RH input units. The activation from this composite input is then fed through to the output. A comparison is then made between the output of the network and the composite input. The maximum of the absolute values of the error across all output nodes is the error measure for item *abc*. In other words, it is a measure of how familiar the chunk *abc* is to the network.

An Example of Chunk Learning in TRACX

Assume that a language contains five words: *abc*, *def*, *ghi, jkl,* and *mno* and that there is a long undifferentiated sequence, *S*, made up of these words:



When the network has arrived at the arrow, assume it has *f* in the LH input units and *a* in the RH inputs. It will do a feedforward–backpropagation pass on this input. Having never encountered this combination of syllables before, the autoassociative error will be high, so on the next time step, it will shift *a* to the LH units and put the next item in the sequence, *b*, in the RH units. But it has encountered *ab* twice before, so the error on output will be lower. If this error is below the error criterion, this tells the network it must have encountered *ab* a number of times before, which means *ab* must be a chunk. So, when the arrow moves to the *c* following *ab*, the network will not put *b* into the LH input vector but rather will put the hidden-unit activation pattern, H_{ab}, which was produced by the autoassociation of *ab* on input to *ab* on output into the LH input vector. Thus, $(H_{ab})c$ is how the network represents *abc* on the input layer. How far will this chunking process go? It will eventually make the chunk H_{abc}, but thereafter, the *c* can be followed by *g*, *m, j, a,* or *d*. Thus, it will not remember that it encountered, say, *abcj*, because it will have encountered that only once for every five times it encountered *abc*. So, unless the string *S* is very long, thereby allowing the network to see *abcj* many times, it will not learn this as a chunk. Further, even if *S* is long, the process of interference from other items will, in general, prevent *abcj* from emerging as a chunk. In other words, low-frequency chunks will simply never be formed.

TRACX and the SRN: Two Fundamentally Contrasting Approaches to Sequence Segmentation and Chunk Extraction

It is important to spell out clearly the major differences between TRACX and its 20-year-old connectionist cousin, the SRN model. In the simulation section, in addition, we present data on which TRACX succeeds and the SRN fails. The SRN learns by predicting the upcoming item in a sequence and by then comparing this prediction with the item that actually appears. Based on the difference between its prediction and the item that really occurs, it changes its weights to bring prediction closer to reality. TRACX relies on a fundamentally different approach to learning; that is, recognizing what it has previously encountered.

One might argue that the context units of an SRN are its way of looking backward. There is some truth to this, but there is a fundamental difference with respect to TRACX. An SRN *always* includes the hidden-unit activations from the previous time step in the input layer, whereas TRACX only does so when it encounters input that it recognizes as having been encountered before. The crucial difference between the two architectures and the philosophy of learning that they implement is that in TRACX, the hidden-unit activations that are put into the input layer are a compressed

representation of the two previous items on input that were recognized as having been encountered before (i.e., an internal representation of a "chunk"). This is not the case in an SRN in which the key notion of the reinjection into the input of internal (compressed) representations of previously encountered subsequences of input is absent.

Since the SRN model is fundamentally tied to prediction, it stands to reason that in situations in which prediction is of no help in producing correct segmentation, it should fail. We show that this is indeed the case.[4]

TRACX Simulations

The TRACX simulations are organized as follows. We first present seven simulations of previously published data, a simulation that predicts data that we subsequently confirmed in an empirical study with infants, a simulation with a large infant-directed language corpus to demonstrate TRACX's ability to scale up and, finally, present two simulations involving bilingual microlanguage acquisition designed to demonstrate two important additional properties of the model—namely, that it automatically clusters the internal representations of the chunks it has discovered and can generalize to new data.

These simulations use data from three areas: infant SL, adult IL, and a real-world infant-directed language corpus. We begin by presenting simulations of two classic experiments by Saffran et al. (1996) and by Aslin et al. (1998). This is followed by five adult IL experiments involving word extraction—namely, two experiments from Perruchet and Desaulty (2008), two from Frank et al. (2010), and one from Giroux and Rey (2009). In all of these experiments, the number of words in the familiarization language was relatively small, ranging from four words in the case of Aslin et al. (1998) to 27 in Perruchet and Desaulty (2008). In order to test the ability of TRACX to scale up, we ran the model on Brent and Cartwright's (1996) corpus of infant-directed speech. The full corpus consists of 9,800 phonetically encoded sentences containing a total of 33,400 words (1,321 different words) and 95,800 phonemes.

In two further simulations, designed to test the generalization and the clustering capabilities of TRACX, we simulated an individual who hears three-syllable words in two separate microlanguages. These simulations are similar to one developed by French (1998) using an SRN in which, over time, the hidden-unit activation patterns (i.e., the network's internal representations) for the words in both microlanguages form two distinct clusters.

We then go beyond existing data by developing an equal TP experiment in which all (forward) TPs, both within words and between words, are equal but in which the backward TP between within-word syllables (1.0) was 4 times that of between-word syllables (.25). TRACX makes a clear prediction about this data—namely, that words will be learned better than partwords—and we show that this prediction is borne out experimentally with 8-month old infants. We also show that the SRN model does not discriminates words from partwords in this simulation as well as TRACX because it can rely only on differential frequency information since all forward TPs are identical.

Modeling Infant SL

Simulation 1: Saffran et al. (1996) Experiment 1

We started with the stimuli from the seminal experiment (Saffran et al., 1996) on infant word extraction. This was the first article to emphasize the importance of TPs in word segmentation of a continuous sound stream. Infants heard a continuous stream of words for 2 min in a technique developed by Jusczyk and Aslin (1995). Six different words were constructed, each composed of three distinct syllables drawn from an alphabet of 12 syllables. A sequence consisting of 90 of these words (270 syllables), randomly selected and with no immediate repeats or pauses between words, was presented twice to 8-month-olds. The infants were then tested to determine whether they were able to discriminate between the words of the language and the partwords consisting of the final syllable from one word and the first two syllables of another word. The forward TPs were higher for syllable pairs within words than for syllable pairs making up partwords. After infants listened to the familiarization sequence, their attention was drawn to a speaker by a flashing red light. Words (or partwords) from the familiarization sequence were repeatedly broadcast from the speaker, and the amount of time that the infant continued to look at the speaker was measured. Infants looked at the speaker significantly longer when they heard partwords, compared with words. Saffran et al. (1996) invoked a novelty preference to explain these results. They reasoned that the infants had learned the words in the language better than they had learned the partwords and, therefore, were more attentive to the more novel partwords, generating longer looking times than for words.

Like Saffran et al. (1996), we created a language of four trisyllablic words and trained TRACX on a random sequence of 180 of these words, with no immediate word repetitions. After familiarization, the network was tested on its ability to discriminate between the words and the partwords made up of the final syllable of one word and the first two syllables of another word. The results are shown in Table 1. Network error is significantly lower for words than partwords. This means that like the infants, TRACX has on average learned the words in the familiarization sequence better than the partwords, thereby allowing it to discriminate between the two.

We also ran an SRN on this task. The SRN we tested had a 2*N*–*N*–*N* architecture (*N* = number of syllables), a learning rate of .01 and momentum .9. We also ran the SRN with learning rates other than .01, but this value gave the best results overall. We tested it in the same way we tested TRACX, that is, by presenting

[4] The SRN architecture could be modified so that part of the network would "predict" syllables that had just been encountered, while the standard part would predict upcoming syllables, and a combined error measure of the two predictions could be used for word/partword discrimination. But this would constitute a fundamental change in the way in which SRNs are conceptualized—namely, as networks that rely on prediction-driven learning. In fact, Maskara and Noetzel (1993) did something related to this by combining an SRN, which predicts the next element, with an autoassociator that recognizes the current element in a sequence. This hybrid was found to improve the natural language recognition ability of a basic SRN. However, it still neither forms chunks nor segments sequences on the basis of backward TPs.

Table 1
Comparison of Humans, TRACX, and an SRN Across Five Studies

Experiment	Simulation	Population	Segmentation cues	Score type	Words learned significantly better than partwords? (Proportion better) Humans	TRACX	SRN
Saffran et al. (1996)	1	Infant	Frequency + Forward TPs	Looking time	Yes (.06)	Yes (.08)	Yes (.68)
Aslin et al. (1998)	2	Infant	Forward TPs	Looking time	Yes (.04)	Yes (.13)	Yes (.58)
Perruchet & Desaulty (2008): Experiment 2	3	Adult	Forward TPs	% Correct responses	Yes (.34)	Yes (.38)	Yes (.80)
Perruchet & Desaulty (2008): Experiment 2	4	Adult	Backward TPs	% Correct responses	Yes (.22)	Yes (.32)	No (−.10)
Equal TP (this article)	9	Infant	Frequency + Backward TPs	Looking time	Yes (.13)	Yes (.50)	Yes (.05)

Note. The values for all simulations are averages from 25 runs of the program. In all cases, the numbers represent how much better words were learned, compared with partwords. See footnote 5 for a detailed explanation of the proportion-better score. TRACX = truncated recursive autoassociative chunk extractor; SRN = simple recurrent network; TP = transitional probability.

bipolar encodings of the test subsequences (words or partwords) to the network, zeroing the context units, and recording the error after the subsequence had been fed through the network. As can be seen in Table 1, the SRN learned words better than partwords on the Saffran et al. (1996) data.[5]

Simulation 2: Aslin et al. (1998) Experiment 1

In Saffran et al.'s (1996) familiarization sequence, in addition to higher TPs for words than for partwords, words were heard 3 times as often as their associated partwords. Aslin et al. (1998) developed an experimental design that removed these frequency effects. They used four trisyllabic words, two of which were low-frequency words and two of which were high-frequency words, the latter occurring twice as often as the former in the familiarization sequence. This meant that the partwords spanning the boundary between the two high-frequency words, which we call HH-partwords, would have the same frequency in the familiarization language as the low-frequency words. At the same time, the internal TPs for the low-frequency words ($TP_{LF-word} = 1$) would be higher than the TPs for the HH-partwords ($TP_{HH-partword} = .5$). During testing, the low-frequency words were tested against the HH-partwords, and it was found that infants looked significantly longer at the HH-partwords, reflecting the better learning of the low-frequency words, compared with the HH-partwords.

After familiarization with a 270-word sequence, as in Aslin et al. (1998), we tested TRACX on low-frequency words and HH-partwords. The model's performance is shown in Table 1. As in Aslin et al., TRACX learned the LF-words significantly better than the HH-partwords. The SRN, with the same parameters as in Simulation 1, also learned words better than partwords for this experiment.

Discussion of Infant Simulations

TRACX successfully models infants' better learning of words over partwords in the context of (a) differential within-word, versus between-word (forward), TPs (Saffran et al., 1996) in an artificial language and (b) differential (forward) TPs with word–partword frequency matching (Aslin et al., 1998). As suggested by Mirman et al. (2010), SRNs were also able to capture these data.

Adult IL

Simulation 3: Perruchet and Desaulty (2008) Experiment 2—Forward TPs

In this and subsequent studies included in this section, adults first listened to a familiarization sequence. They were then given a series of single two-alternative forced-choice tests between a randomly chosen word and a partword, both of which had occurred in the familiarization sequence, and were asked which they were most confident they had heard during familiarization.

In this simulation, we used the syllable sequences constructed by Perruchet and Desaulty (2008), with a localist encoding scheme as in the infant simulations. Nine different two-syllable words were constructed from a set of 12 syllables: *a, b, c, d, e, f, g, h, i, x, y,* and *z.* These words were concatenated, with no breaks between them, into a familiarization string 1,035 words long. All

[5] In Table 1, we use a "proportion better" measure to compare model results and empirical data. This is a relative-difference measure and can be applied equally well to error measures or to looking times. So, for example, in Saffran et al. (1996) Experiment 2, babies looked at words for 6.77 s and partwords for 7.60 s. This means that there was an absolute difference in looking times of 0.83 s. The larger this difference, the better words have been learned. Obviously, however, if the initial looking-time values had been, say, 12.0 s and 12.83 s, this would also have given a difference of 0.83 s, but the relative difference would have been much smaller. For this reason, we need to "normalize" the absolute difference by the total looking time at both items. This gives 0.83/(6.77 + 7.60) = 0.06, which indicates how much better words have been learned relative to partwords. This is the measure we call the "proportion better" of word learning, compared with partword learning. We use the same measure for the output error of the model. Since the lower the error, the better the learning, we can compare the difference between the output error for words (1.41) and the output error for partwords (1.65) on the Saffran et al. (1996) familiarization data and conclude that TRACX learned words better than partwords, the absolute output-error difference being 0.24. As with looking time scores, we need to normalize this value by the total error score (1.41 + 1.65) to get the relative amount of how much better words were learned by TRACX, compared with partwords—in this case, 0.24/(1.41 + 1.65) = 0.08. In short, the proportion-better measure provides a simple means of comparing empirical data with data produced by our model.

words were frequency-balanced, occurring 115 times each in the familiarization sequence. The internal forward TP between syllables of all words was 1. So, for example, one of the words was *xa*, which meant that whenever *x* occurred, it was followed, with a probability of 1, by an *a*. Any of the other 11 letters could precede *a*, but *x* was always followed by an *a*. Words are learned significantly better than partwords (Table 1).

As with the adults in this study, TRACX also learns the words significantly better than the partwords. We also tested the SRN model (with the same parameters as in the previous simulations). It, too, was able to learn words better than partwords in this condition. All results are shown in Table 1.

Simulation 4: Perruchet and Desaulty (2008) Experiment 2—Backward TPs

Perruchet and Desaulty (2008) were the first to explicitly show that *backward* TPs can be used by adults' as a cue for word-extraction. This result constitutes a critical experiment for the work reported here.

The experimental methodology was identical to that described in Simulation 3, except that backward TPs were manipulated. So, for example, in the present study, *XA* is a 2-syllable word made up of the syllables *X* and *A*. This word has an intrasyllable backward TP of 1, which means that *A* can only be preceded by *X*. However, *X* can be followed by any of the syllables in the set {*A*, *B*, *C*}, where *A* occurs 3 times as often as *B* or *C*. The frequencies of the words in the familiarization sequence were adjusted so that as in Aslin et al. (1998), the frequencies of the test words and partwords in the familiarization sequence were identical. Forward TPs were higher between the syllables on either side of a word boundary (.33) than between the syllables within a word (.20). Based on FTP information, this would have meant that partwords would be better recognized than would words. However, the backward TPs within a word (i.e., for the word *XA*, the probability of *X*, given *A*, written $p[X|A]$ is 1.0) were considerably higher than the backward TPs between words (.20). Thus, if participants segmented at word boundaries rather than at partword boundaries, one would have to conclude that they had relied on backward TP cues to do so.

Participants in a forced-choice recognition test chose words significantly more often than partwords. TRACX recognizes words better than partwords. However, when we tested an SRN on this task (with the same parameters used in the other SRN simulations), it chooses partwords more often than words. All results are shown in Table 1.

Simulation 5: Giroux and Rey (2009)

Giroux and Rey (2009) showed that sublexical units (which we call subchunks) found within words become more difficult to identify once they have been combined into a larger chunk. They compared the recognition of the subchunks making up a word early in learning and late in learning of the word. They found that recognition performance for subchunks incorporated into larger chunks decreased as learning proceeded. We simulated Giroux & Rey as follows. We created a familiarization sequence made up of two-, three- and four-syllable words. We then exposed TRACX to this sequence, repeated 30 times. We tested how well various subchunks making up the word were recognized by the network at various points during learning.

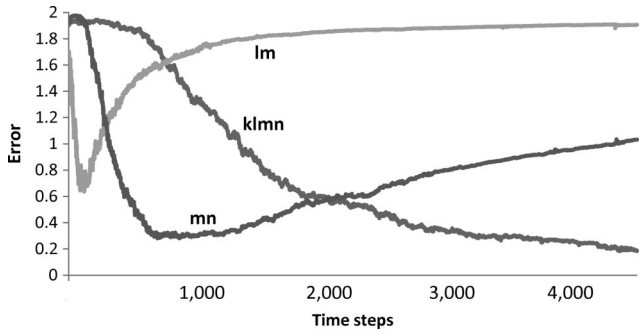


Figure 2. The subchunks *lm* and *mn* are initially extracted as words by TRACX, but as soon as the system recognizes the larger chunk, *klm*, the subchunk *lm* will gradually be forgotten by the network if it only appears inside the chunk *klm*. Later, when the system recognizes that *klmn* is a chunk, *mn* will also cease to be seen as a separate entity, and the network will gradually forget it.

We tested TRACX's learning of the 4-syllable word *klmn* and considered its subchunks *kl*, *lm*, and *mn*. Initially, the program detects separate chunks *kl*, *lm*, and *mn* and begins to learn these chunks. The error on these three small chunks drops rapidly. However, *kl* and *lm* soon become a single chunk, *klm*, because *m* is chunked to *kl*. Thereafter, *lm* will no longer be chunked separately because it always appears as part of *klm*. Thus, while the recognition error for *lm* initially dropped rapidly (when the network still considers it to be an independent chunk), when *lm* becomes incorporated into *klm*, the error associated with *lm* rises (see Figure 2). Once *klm* is formed and systematically recognized by the network, *n* begins to be chunked to *klm* to form *klmn*. Thereafter, the subchunk *mn* gradually begins to be forgotten by the network, and the recognition error for *mn*, like *lm* before it, begins to increase (see Figure 2). Thus, for these two internal subchunks (i.e., *lm* and *mn*), TRACX replicates the results of Giroux and Rey (2009).[6,7] This is presumably what happened in English to words like *mand* and *monish* that were once independent words but, through gradual lack of refreshment through use, today survive only as subchunks of words like *reprimand* and *admonish*.

This addresses a deep issue about chunking—namely, that all chunks are not equally atomic. Chunks tend to become more

[6] As for the leading chunk *kl*, TRACX continues to believe it is a separate chunk. This problem would, presumably, disappear in the more advanced version of TRACX referred to in the final paragraph of the TRACX Implementation section, in which "lookahead" would spot previously discovered chunks in the raw input and would allow the network's internal representations of these chunks to be put into the RH-side of the input.

[7] We also tested the SRN on the Giroux and Rey (2009) data for various sizes of the hidden layer. In contrast to TRACX, we were unable to get the SRN to reproduce the empirically observed "early learning followed by forgetting" rebound (see Figure 2) that occurs when subchunks gradually become internalized within larger chunks (Giroux & Rey, 2009; Perlman et al., 2010).

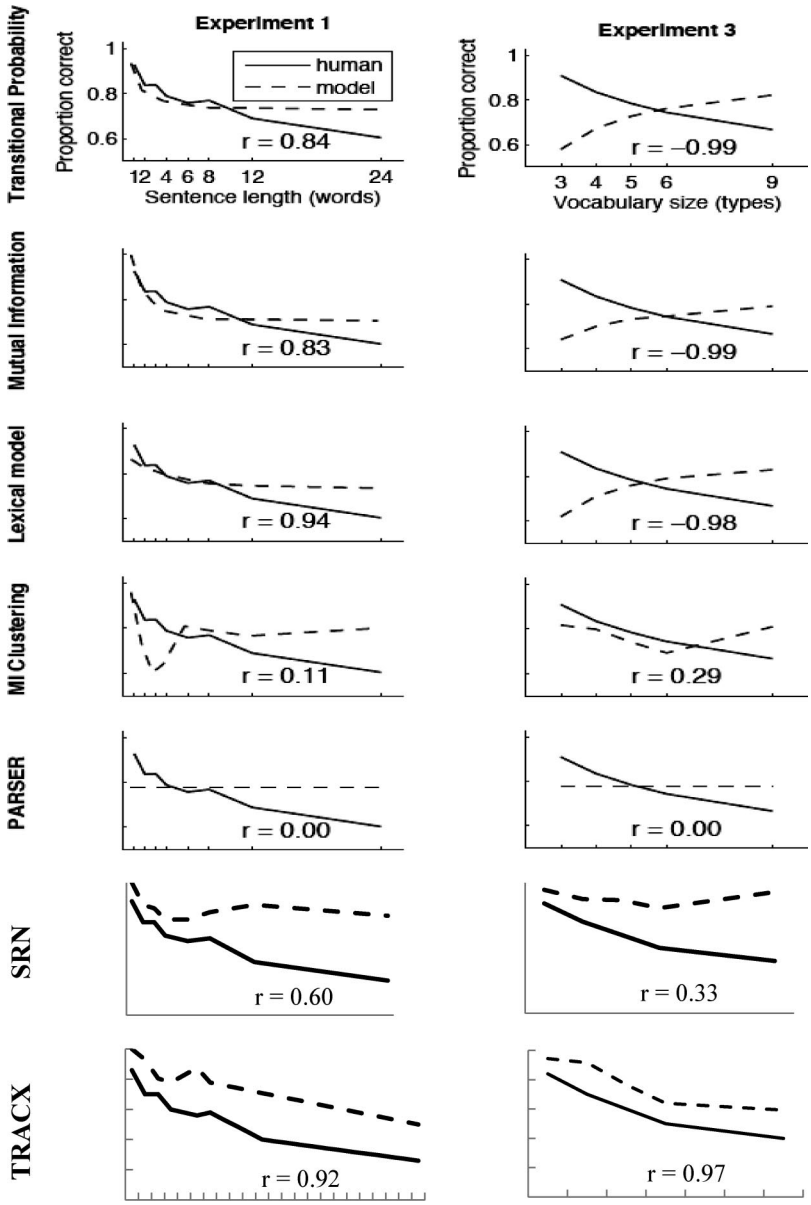


Figure 3. A comparison of the performance of seven models of adult implicit learning on two segmentations tasks as a function of sentence length (left column) and vocabulary size (right column). Solid lines represent human data, and dashed lines represent model data. All non-TRACX and non-SRN data are from "Modeling Human Performance in Statistical Word Segmentation," by M. C. Frank, S. Goldwater, T. L. Griffiths, and J. B. Tenenbaum, 2010, *Cognition, 117,* p. 115. Copyright 2010 by Elsevier. Reprinted with permission. SRN = simple recurrent network; MI = mutual information.

atomic with use, and their subchunks become gradually harder to notice. Compare *breakfast* and *nosebleed*, *congressman* and *congresswoman*, and *cupboard* and *fleabite*. In each of these three pairs of words, we hear the subchunks more clearly in the second word. It is for this reason that it is important in later models of TRACX to make the mechanism by which chunks are inserted into the input layer stochastic rather than deterministic, the probability of insertion being a function of the error on output on the previous time step.

Simulation 6: Frank et al. (2010) Experiment 1

Frank et al. (2010) observed that as sentence length increased, it became more difficult for adult participants to extract the words from the familiarization language. They used a corpus of 18 syllables with which they created two 2-syllable words, two 3-syllable words, and two 4-syllable words. No words shared syllables. They presented a continuous sound stream to participants made up of these words in which there was a brief pause

after each sentence. They tested participants in eight conditions in which sentences contained one, two, three, four, six, eight, 12, or 24 words. To simulate this, we created two 2-, two 3- and two 4-syllable words (*ab*, *cd*, *efg, hij, klmn,* and *opqr*) and used them to build a set of eight different sequences of 144 words, containing 144, 72, 48, 36, 24, 18, 12, and six sentences. The learning procedure was as before. Each sentence was presented to TRACX six times during the learning phase. Figure 3 (left panels) shows the difference between TRACX's errors for words and partwords, that is, the means by which TRACX discriminates between words and partwords. The larger this value, the better the discrimination. This difference drops as sentence length increases, meaning that the longer the sentences in which the words and partwords are found, the harder it is for TRACX to discriminate between them. This is in agreement with the results of Frank et al. (2010), who found that the percentage of words correctly extracted from texts fell off as the sentences in which the words were found grew longer. PARSER was run on these data, and its output did not correlate with the human data found by Frank et al. ($r = 0$). The SRN does well on this task (averaged over 25 runs with 12 different sequences per run), producing a correlation to human data of approximately $r = .60$. TRACX, by comparison, produced output whose correlation with human data is $r = .92$ (see Figure 3, bottom left panel).

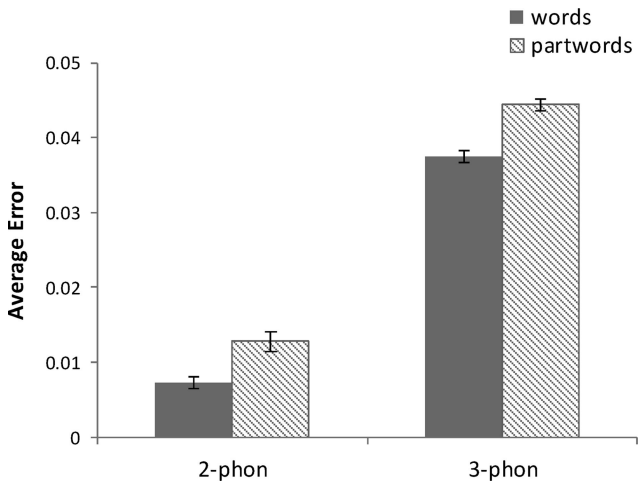


Figure 4. TRACX's differential word and partword output errors allow it to distinguish two- and three-phoneme (phon) words from partwords in the full corpus. Values plotted are average absolute error across all output units (i.e., $\frac{1}{2N}\sum_{i=1}^{2N} \mid \text{out}_i - \text{in}_i \mid$, where $2N$ is the size of the input and output layers and N is the number of phonemes), and error bars show standard errors of the mean.

Simulation 7: Frank et al. (2010) Experiment 3

The more words there are in a language, the more difficult it should be to learn (and remember) the vocabulary of that language. Frank et al. (2010) tested this hypothesis by varying the number of word tokens in an artificial language and observing how well participants can distinguish the words of the language from partwords.

All sentences contained four words, and the pool from which the words making up each sentence was drawn consisted of between three and nine different words, depending on the condition. Words varied in length from two to four syllables. There were equal numbers of two-, three- and four-syllable words in the familiarization text.

This experiment is particularly important for the present article because none of the five major models examined by Frank et al. (2010)—namely, the TP model (Saffran et al., 1996), the pointwise mutual information model (Frank et al., 2010), the Bayesian lexical model (Goldwater, Griffiths, & Johnson, 2006, 2009), the mutual information (MI) clustering model (Swingley, 2005), and PARSER (Perruchet & Vintner, 1998, 2002)—succeeded in capturing the human pattern of performance (Frank et al., 2010, p. 116). The correlations of the percentage of correct outputs of each of these models with human data were −.99, −.99, −.98, .29, and .00, respectively (see Figure 3, right panels). By means of a somewhat counterintuitive reduction of the amount of data presented to the models to a mere 4% of the original amount, Frank et al. were able to considerably improve the performance of the TP model and the lexical model (cf. Figure 4 in Frank et al., 2010). PARSER's correlation to human data on this task was $r = 0$. The simulations we ran with the SRN (averaged over 25 runs with 12 different sequences per run) showed a correlation of the output to human data to be $r = .33$. With the same parameters as for the previous simulations, TRACX produces output whose correlation with human data is $r = .97$ (see Figure 3, bottom right panel).

Discussion of Adult Studies

TRACX successfully models adults' better learning of words over partwords in the context of (a) differential within-word, versus between-word, forward TPs (Perruchet & Desaulty, 2008) in an artificial language with frequency-controlled test words and partwords; (b) differential within-word, versus between-word, backward TPs (Perruchet & Desaulty, 2008) in an artificial language with frequency-controlled test words and partwords; (c) gradual forgetting of subchunks found inside chunks (Giroux & Rey, 2009), if these subchunks are not independently refreshed; (d) sentence length and the fact that words become harder to extract as the length of the sentences in which they are found increases (Frank et al., 2010); and (e) vocabulary size and the fact that words become harder to extract as the number and length of the words to be extracted increases (Frank et al., 2010).

Moreover, in contrast to TRACX, prediction-based SRNs were unable to extract words better than partwords in the case in which backward TPs were the cue for chunking. In addition, for the experiments with increased sentence length and increased vocabulary size (Frank et al., 2010), PARSER's output was uncorrelated to human data in both cases, whereas the output from TRACX was highly correlated to human data.

Scaling up to the Real World and Generalization

In Simulations 8, 9, and 10, we explore two crucial issues. The first is to see how well TRACX can scale up to the real world. The second examines the organization of the distributed internal representations of TRACX. We also show that this distributed orga-

nization allows TRACX to generalize in a completely natural manner to input that it has never previously encountered.

Simulation 8: Scaling up to the Real World—Brent and Cartwright (1996)

For any cognitive model, it is crucial to show that the model can at least potentially scale up to phenomena in the real world. To explore TRACX's ability to scale up, we tested it on a real-world infant-directed language corpus prepared by Brent and Cartwright (1996), using data from Bernstein-Ratner (1987). The first 10 sentences from this corpus, along with their phonetic encodings, are shown in Table 2.

The Brent and Cartwright (1996) corpus was several orders of magnitude larger than all of the other corpora used in the other simulations. For example, in the Aslin et al. (1998) study, a total of four different words were used, and in Saffran et al. (1996), 6 different words were used. By contrast, the Brent and Cartwright corpus consisted of 1,321 separate words, 9,800 sentences containing a total of 33,400 words, and over 95,800 phonemes. This meant that there were often long separations between repetitions of the same word. A .04 learning rate was used in all of the previous simulations with a small number of words and small corpora to ensure rapid learning. However, in the very large Brent and Cartwright corpus, repetitions of the same word were often very widely spaced, which meant that by the time a new occurrence of a particular word occurred, the learning rate of .04 would have changed the weights enough to have effectively erased all traces of earlier occurrences of the word. We, therefore, lowered this learning rate to .005 for the present simulation. A sentence was presented to the network six times before moving on to the next sentence. The architecture of the network and all other parameters were identical to the other simulations. TRACX was trained on five passes through the training corpus. It should be recalled that only 12 of the 112 two-phoneme words—that is, *yu, D6, Iz, It, tu, du, D*, si, In, y), y&*—and seven of the 384 three-phoneme words—that is, *D&t, WAt, DIs, lUk, k&n, &nd, oke*—in the corpus occur more than 1% of the time. In other words, TRACX is learning these small phoneme-chunks, even though it has very little overall exposure to them.

Table 2
The First 10 Sentences From the Brent and Cartwright (1996) Child-Directed Language Corpus

English	Unsegmented phonemes
You want to see the book	*yuwanttusiD6bUk*
Look there's a boy with his hat	*lUkD*z6b7wIThIzh&*
And a doggie	*&nd6dOgi*
You want to look at this	*yuwanttulUk&tDIs*
Look at this	*lUk&tDIs*
Have a drink	*h&v6drINk*
Okay now	*okenQ*
What's this	*WAtsDIs*
What's that	*WAtsD&t*
What is it	*WAtIzIt*

Note. From "Distributional Regularity and Phonotactic Constraints Are Useful for Segmentation," by M. R. Brent and T. Cartwright, 1996, *Cognition, 61,* 93–125. Copyright 1996 by Elsevier.

The testing procedure was patterned after the one used in the other experiments, in which the experimenters tested participants on words and partwords from the training corpus. We took a number of two- and three-phoneme partwords from the text and tested them against two- and three-phoneme words in the text. Because of the large number of phonemes (90), we used average error (rather than maximum error) over all output units as the measure of error. This was because a maximum-error-across-all-outputs measure does not distinguish between almost-perfect learning (e.g., in which only one output unit is wrong) and terrible learning (e.g., in which most of the output units are wrong). An average-error-across-all-outputs measure solves this problem. The results of a run of TRACX on this database are shown in Figure 4. Even with this much larger corpus, words are learned better than partwords.

Simulations 9 and 10: Generalization and the Organization of TRACX's Internal Representations

In order to demonstrate the ability of TRACX to develop hidden-unit activation clusters that reflect structure in the world and to show how it can generalize to new input, we have chosen an example based on work using an SRN to model bilingual language acquisition (French, 1998). We ran two simulations, one in which it is relatively easy to demonstrate the methodology and one that is considerably more demanding, to show that TRACX really can extract structure from its input.

For the easy simulation, two microlanguages, Alpha and Beta, consisting of three-syllable words were created. Each language had its own syllable sets: initial, middle, and final syllables. The Alpha language consisted of initial syllables {*a, b, c*}, middle syllables {*d, e, f*}, and final syllables {*g, h, i*}. In like manner, the Beta language consisted of initial syllables {*j, k, l*}, middle syllables {*m, n, o*}, and final syllables {*p, q, r*}. A word in a given language consisted of a random initial syllable, followed by a randomly chosen middle syllable, followed by a randomly chosen final syllable, all chosen from the syllable set corresponding to that language. Thus, *beh* was a word from Alpha; *knp* was a word from Beta. There were no markers indicating either word boundaries or language boundaries. A typical language familiarization sequence of syllables might look like this: *adgbehbdgcficdgafglnrloplmpknpjmrkmpbfgbehcdiafh*. We generated a total of 10,000 words, approximately 5,000 from each language. These words were drawn from a subset of two-thirds of the possible words from each language. At any point, there was a probability of switching to the other language after each word of $p = .025$. TRACX, whose parameter settings were identical to those used in Simulation 8 with a learning rate of .001, was given this familiarization sequence. After one pass through the familiarization corpus, the network was tested on the full set of possible words in both languages (i.e., including the one-third of possible words in each language that it had never seen). For each possible word in each language, we recorded the hidden-unit representation that was produced. We then did a standard cluster analysis (Ward's method, Euclidean distance between vectors) and were able to observe that the network consistently produced two distinct clusters corresponding to the words of each language. In addition, with very few errors, all of the previously unseen items

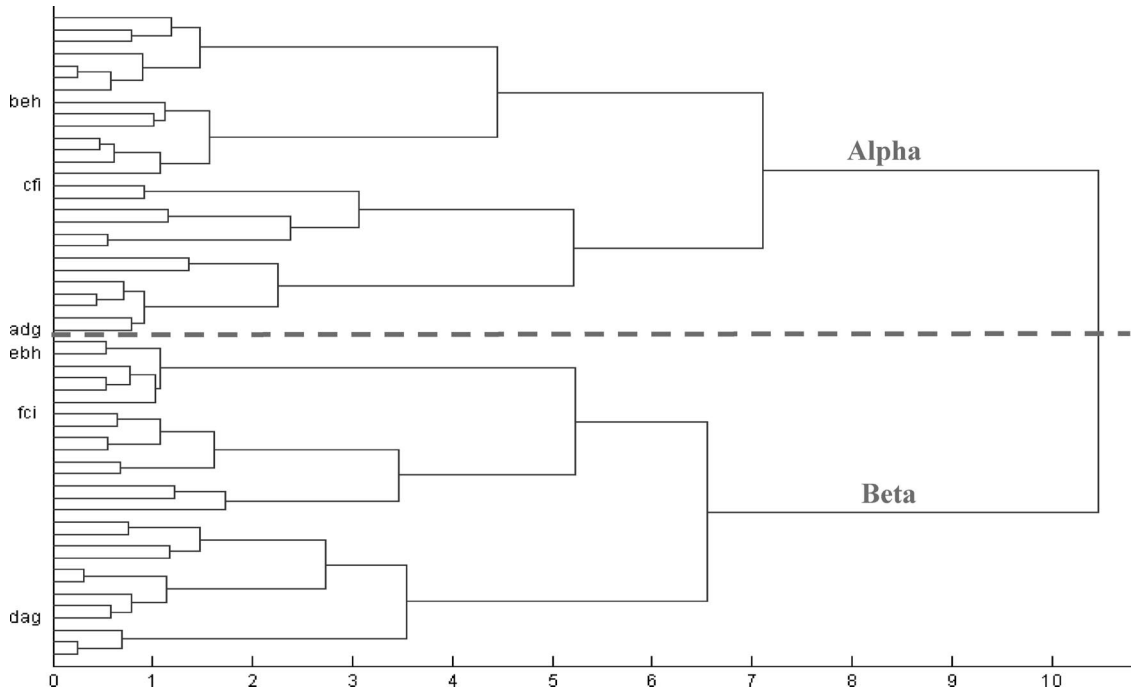


Figure 5. Dendrogram of TRACX's internal representations of three-syllable words from two microlanguages, Alpha and Beta, after training on a corpus containing sequences of words from both languages. Alpha and Beta use the same set of nine syllables to create their words and differ only on the structure of the words. TRACX's internal representations of the words from the two languages form two separate clusters, as shown in the figure. TRACX was also tested on six new words that never occurred in the training sequence. Half of the test words (*adg, beh, cfi*) had the structure of Alpha words. The other half (*dag, ebh, fci*) had the structure of Beta words. All six words were correctly classified.

from each language (i.e., one-third of all possible words from each language) were categorized correctly.

However, this simulation left open the question of whether TRACX was sensitive to the structure of the words or was simply basing its classification on the nonoverlapping syllable sets for each language. We, therefore, ran a second simulation in which the Alpha language was the same as before but in which Beta used the same syllable set as Alpha. However, the initial syllables of words in Beta were the middle syllables of words in Alpha, and the middle syllables of words in Beta were the initial syllables of words in Alpha. In other words, for *adg* in Alpha, there was a corresponding word, *dag*, in Beta. A bilingual language stream of 10,000 words was produced as before, with a switching rate of .025. Six words, three from Alpha and three from Beta, were left out of the training sequence. When we analyzed the hidden-unit representations as before, we found that TRACX was capable of correctly clustering the languages and of correctly classifying the six unseen words (Figure 5).[8]

Summary of Scaling up, Generalization, and Clustering of Internal Representations

TRACX is able to scale up to corpora far larger than those in use in typical experiments involving infant SL and adult IL. TRACX also develops internal representations that reflect the structure in the environment and can generalize appropriately based on that structure.

Beyond Existing Data

Simulation 11: Box Languages—Flexible Use of Frequency and TP Information

Simulations 1 through 10 show that TRACX can use backward and forward TP information to identify word boundaries in a stream of input. However, experimentally, it is difficult to fully control word/partword frequency and TPs simultaneously. For example, if forward TPs are balanced then, in general, the frequencies of words and partwords in the text and/or backward TPs are not.

[8] The network does not always classify the hidden-unit representations for words in each language (a total of 54 words) into exactly two perfect clusters, as shown in Figure 5, but generally forms a small number of large subclusters of representations of words from each language. This is, in general, sufficient for correct classification of novel items. The exact conditions (learning rate, chunking criterion, number of words, etc.) required to always produce perfect clustering for overlapping languages of this type is a question for future research. The point is that TRACX is capable of discriminating the two languages based on their structure alone.

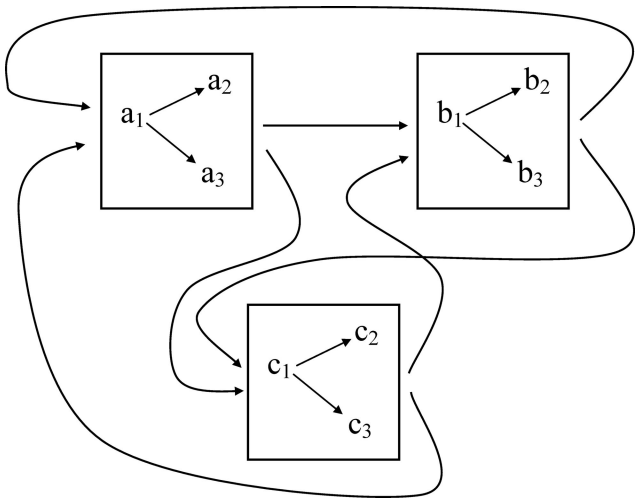


Figure 6. The three "word" blocks A, B, and C. Between syllables within a word (a_1, a_2, a_3; b_1, b_2, b_3; and c_1, c_2, c_3), the forward transitional probabilities (TPs) are all .5, and between-word forward TPs are also .5.

We are aware of only two experiments to date that directly examined the extent to which backward TPs can be exploited for word extraction from a syllable stream. In the first, Perruchet & Desaulty, 2008, explored whether adults could use backward TPs to do word extraction; the second was a very similar study with infants, in which Pelucchi et al., 2009a, used a natural language corpus of real Italian words.

In both of these experiments, the frequencies of various words found in the familiarization sequence but absent from the test phase ensure that the test words and partwords have certain TPs. For example, in the case of Perruchet and Desaulty (2008, Experiment 2), the frequency of the words *XA*, *YD*, and *GZ* in the familiarization sequence is 3 times as high as the frequency of the words and partwords *AX* and *XB* that were presented during the test phase. The assumption, potentially problematic, is that this frequency imbalance has no influence on the words and partwords tested. We therefore designed an experiment in which we controlled for forward TPs and, at the same time, ensured that all words (i.e., *A*-words, *B*-words, and *C*-words, see Figure 6) in the familiarization sequence had the same frequency of occurrence.

The construction of the box language is shown in Figure 6. There are three word blocks: A, B, and C. Between syllables within a word, the TPs are all .5. So, $p(a_2|a_1) = .5$, $p(a_3|a_1) = .5$, $p(b_2|b_1) = .5$, $p(b_3|b_1) = .5$, $p(c_2|c_1) = .5$, $p(c_3|c_1) = .5$. The TPs between words are also .5. Thus, $p(A|B) = .5$, $p(B|C) = .5$, $p(C|A) = .5$, and so on. This means that every *A*-word, say, a_1a_2, can be followed by either a *B*-word or a *C*-word, with equal probability. So, a_1a_2 can be followed by b_1b_2, b_1b_3, c_1c_2, or c_1c_3. In other words, a_1a_2 will be followed half the time by b_1 and half the time by c_1. Thus, the word a_1a_2 will occur twice as often as the partwords a_2b_1 and a_2c_1. This means that forward TPs are identical for all the syllable pairs, but word frequencies (i.e., frequencies of *A*-words, *B*-words, and *C*-words) are twice *partword* frequencies. This construction implies that backward TPs within words (=1) are higher than those between words (=.25).

We first ran TRACX on a familiarization sequence of 135 words (270 syllables) produced from this design and then tested the network in the usual way on words and partwords from the sequence. As can be seen in Figure 7 (left panel), TRACX learns the words significantly better than it learns the partwords. It does not need forward TP information to do chunk extraction and can flexibly adapt the basis of its computation, depending on the available informative cues. We also ran an SRN on the same familiarization sequence. As shown in Table 1, the SRN does not distinguish between words and partwords particularly well in this simulation.

By simply "flipping" the transitions in each box (i.e., the *A*-words would become a_1a_3 and a_2a_3, the *B*-words would become b_1b_3 and b_2b_3, and the *C*-words would become c_1c_3 and c_2c_3), this would produce a dual box language in which all backward TPs are .5, but forward TPs were 1 for words and .5 for partwords, with partword frequencies being twice that of word frequencies. We suggest that simulations and experimental designs of this sort might be used to tease out the relative contribution of frequency and TP cues in sequence segmentation.

Testing a Model Prediction: Infant Segmentation With Equal Forward TPs

TRACX can rely equally well on forward and backward TPs and on frequency information to extract words from an input stream depending on cue informativeness. Here, we establish that 8-month-olds can also do so. Having first replicated the key Aslin et al. (1998) results (albeit with a different regional accent),[9] we generated new training and test stimuli that matched the grammar described in Simulation 11, above. We tested a group of infants using a preferential head turn procedure identical to that in the original Aslin et al. (1998) work.

Method

Participants. The participants were 20 eight-month-old infants (12 female, eight male; age $M = 251$ days; range: 226–279 days). A further 11 were excluded due to fussiness (nine), equipment failure (one), or parental interference (one).

Stimuli. New stimuli were generated with the MBROLA (Dutoit, Pagel, Pierret, Bataille, & Van der Vrecken, 1996) speech synthesis package, using a British English voice. Samples were recorded at a bit rate of 16 kHz. Two continuous familiarization sequences, each lasting 3 min (810 syllables at 4.5 syllables/s), were generated using the words shown in Table 3. The test stimuli consisted of the six words and of six of the partwords, each

[9] In order to (a) validate the procedure as run in our laboratory and (b) ensure that the Aslin et al. (1998) findings would generalize to utterances pronounced in British English, we first ran an exact replication of the original Aslin et al. (1998) study showing that 8-month-olds could use forward TP information to extract words from a speech stream. Thirteen 8-month-olds (six girls, seven boys; *M* age = 249 days, range: 232–277) took part. The method and procedures are described in detail in the original article and in the experiment reported in the main text of the current article, with the exception that words were pronounced in British English. As in the Aslin et al. (1998) study, infants looked significantly longer, $t(12) = 2.57$, $p < .025$, two tailed, at partwords ($M = 12.6$ s, $SD = 3.67$) than at words ($M = 10.7$ s, $SD = 3.72$), confirming that infants can use forward TPs to identify word boundaries when frequencies are equal.

Table 3
The Words and Partwords Used in the Equal TP Simulation and Experiment

	Grammar 1			Grammar 2		
Word block	Words	Partwords		Words	Partwords	
A	fee-go fee-rou	go-nei rou-nei	go-duh rou-duh	tai-lu tai-duh	lu-koi duh-koi	lu-rou duh-rou
B	nei-pau nei-koi	pau-fee koi-fee	pau-duh koi-duh	koi-pau koi-nei	pau-tai nei-tai	pau-rou nei-rou
C	duh-lu duh-tai	lu-fee tai-fee	lu-nei tai-nei	rou-go rou-fee	go-tai fee-tai	go-koi fee-koi

Note. The six items in each Word column were used to generate a continuous familiarization sequence with the equal transition probabilities, according to the grammar specified in the description of the equal TPs experiment. In the test phase, these six words were compared to six of the corresponding partwords. TP = transitional probability.

presented on its own, with a 500 ms gap between them, and repeated until the infant looked away.

Procedure. The procedure matched exactly the methods reported in Aslin et al. (1998). Infants were tested in a darkened, soundproofed booth, seated on their caregiver's lap. Throughout the experiment, the caregiver wore headphones and listened to music that masked the sounds presented. Two side-mounted speakers were equidistant from the infant and each had a LED lamp mounted on top of it. A third lamp was mounted directly in front of the infant above a low-light video camera that recorded the infant's behavior. The stimuli were played back on Bose Companion 2 stereo speakers, controlled by a PowerMac G4 computer running Matlab R2007b software. The computer also controlled the blinking lights by means of a Velleman k8055 universal serial bus (USB) interface board. An experimenter, seated in a separate room, observed the infant and controlled the computer.

During familiarization, the 3-min sequence was played continuously over both speakers. The lamp above one of the speakers was always blinking on and off. But the illuminated side changed at random intervals throughout familiarization. The test phase consisted of 12 randomly ordered test trials, such that each test item was presented three times. A trial began with the experimenter illuminating the central lamp to attract the infant's attention. When the infant was looking forward, the central lamp extinguished, and one of the side lamps started flashing. When the observer determined that the infant had looked in the appropriate direction (a 30° head turn), he pressed a key, and the computer played the test word. The word played repeatedly with a 500 ms silent interval between repetitions. A trial ended when the infant looked away from the stimulus for at least 2 s. The experimenter was blind to the order of the trial types.

Results. Infants looked significantly longer at partwords ($M = 12.2$ s, $SD = 2.92$) than at words ($M = 9.4$ s, $SD = 2.89$), $t(19) = 4.62$, $p < .001$, two tailed (see Figure 7, right panel).

Discussion of Infant Experiment

The familiarization sequence in this experiment, based on the word/partword discrimination paradigm developed by Saffran et al. (1996) and Aslin et al. (1998), was designed in such a way that all forward TPs, both between syllables and between words, were identical (i.e., .5). However, even though all forward TPs were equal, the frequency of words was nonetheless twice that of partwords; thus, *backward* TPs were different for words and part-

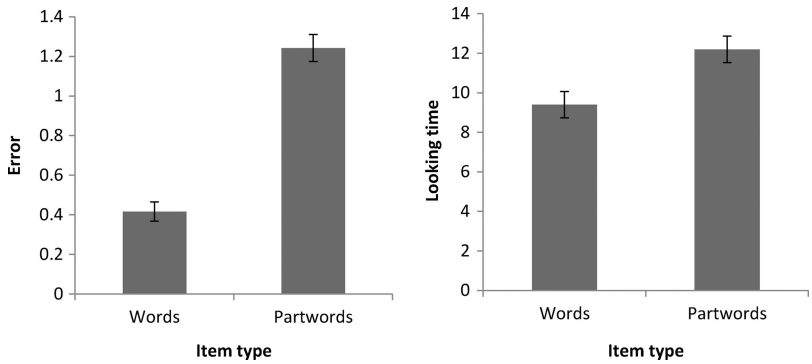


Figure 7. Left: TRACX learns the words significantly better than the partwords, even when all forward TPs (both between syllables within words and between syllables on either side of word boundaries) are identical, but word frequencies are twice as high as the associated partword frequencies. Results are averaged over 25 runs of the program. Error bars are standard error of the mean. Right: Infants learn words better than partwords, even when all TPs are equal, but word frequencies are twice as high as associated partword frequencies ($n = 20$). Error bars are standard error of the mean. TPs = transitional probabilities.

words. Infants could therefore rely on word/partword frequency cues or backward TPs, but not on forward TPs, to extract words from the speech stream. As Figure 7 shows, they were able to discriminate words from partwords, looking time for words being significantly less than that for partwords. We assume, along with Saffran et al. (1996) and Aslin et al. (1998), that a novelty preference was the cause of the longer looking times at the partwords. In other words, the more frequently encountered words had been learned better than the less frequently encountered partwords.

It would, therefore, seem that infants—like the TRACX model—base word extraction on whatever cues happens to be available to them, including forward or backward TPs as well as syllable or word frequencies, in order to ascertain whether they have previously heard a particular sequence of syllables.

General Discussion

In this article, we present TRACX, a connectionist model of sequence processing and chunk extraction based on a mechanism that we have called implicit chunk recognition (ICR). ICR relies on the recognition of previously (and frequently) encountered subsequences of patterns (chunks) in the sequence being processed. This model is intended to provide a unifying framework for the ubiquitous cognitive skills of sequence processing and chunk extraction. Unlike symbolic/hybrid models of chunk extraction (e.g., PARSER; Perruchet & Vintner, 1998, 2002), it requires no WM in which chunks are explicitly stored (and deleted) and does not carry out operations on the items in WM. It also stands in contrast to prediction-based connectionist models of sequence processing, in particular, the SRN (Elman, 1990; Cleeremans & McClelland, 1991) and the TRN (Dominey et al., 2000). Moreover, TRACX successfully captures adult human data when four different normative statistical models (and PARSER) failed to do so (Frank et al., 2010). TRACX was also able to successfully simulate empirical results from the infant SL literature, among them, classic results from Saffran et al. (1996) and Aslin et al. (1998), as well as results from real-world corpus data on infant-directed speech (Brent & Cartwright, 1996).

We ran a total of nine simulations based on existing empirical data and a simulation of bilingual microlanguage acquisition that was designed to illustrate TRACX's ability to generalize and acquire clusters of internal representations reflecting categories in the world. The success of TRACX at simulating a wide range of results suggests that the model can provide new insights into questions involving sequence processing and chunk extraction, in general, and infant SL and adult IL, in particular. Specifically, it suggests that (a) chunk formation proceeds via a process of recognition of previously encountered groups of items rather than via a process of prediction-driven learning and that (b) sequence processing and chunk extraction are graded, rather than all-or-none.

As with all connectionist systems, chunk knowledge is implicit in TRACX. Indeed, knowledge in connectionist systems is described as being in the system, in the sense that it can be used by the system but is not accessible *to* the system, in the sense that it cannot be immediately redescribed by the system in a more abstract or explicit format. In this sense, any chunks extracted from a sequence of inputs are implicit.

In contrast to PARSER (Perruchet & Vintner, 1998), which uses a perceptual-shaping threshold to cull from WM all chunks whose strength fall below this threshold, TRACX has a recognition threshold, which is an error threshold, below which the items on input are considered to form a chunk. The crucial difference with PARSER's perceptual-shaping threshold is that no information is explicitly removed from the TRACX system based on this chunk-recognition threshold. The lack of appeal to explicit processes makes TRACX far more suitable for modeling learning in prelinguistic infants and is consistent with the finding that adults often improve their performance in AGL tasks but are unable to provide verbal reports of the underlying syntactic structures in the task (Reber, 1967). This is not to say that explicit knowledge plays no role in adult performance, but the relation between implicit and explicit learning is a complex one (e.g., Dienes & Perner, 1999; French & Cleeremans, 2002). The issue of how they could both be implemented in a single system and how they would interact is beyond the scope of this article but has been discussed extensively elsewhere (see, for example, Pothos, 2007, for a careful discussion of these issues in the context of AGL).

Word Learning Versus Lexical Segmentation

TRACX performs sequence segmentation and chunk recognition. It was tested on segmentation in the dual domains of adult IL and infant SL. However, it does not connect the chunks it has found to concepts in the world. Rather, it is doing the precursor to word learning, which is lexical segmentation. Word learning is about learning the reference between a symbol (usually an auditory symbol) and a referent in the world (Bloom, 1997; Plunkett, Sinha, Moller, & Strandsby, 1992; Quine, 1960; Waxman & Booth, 2001). Lexical segmentation is assumed to be a prerequisite of word learning in that it is the process of deriving what the informative sound symbols are in the infant's language environment. Only once these have been identified can the infant begin to learn the relevant semantic mappings between the "word" and its referent, and indeed, the prior segmentation of words in a sound stream facilitates word learning both in adults (Mirman, Magnus, Graf-Estes, & Dixon, 2008) and infants (Graf-Estes, Evans, Alibali, & Saffran, 2007; see also Mirman et al., 2010). In the adult AGL literature, all semantic content is removed from the test stimuli specifically to get at this presemantic chunking mechanism. It is, therefore, worth noting once again how much the IL and SL domains share common goals and assumptions, suggesting that a common mechanism may be operating in both domains.

Poverty of the Stimulus Revisited

TRACX allows us to address some of the so-called "poverty of the stimulus" criticisms of statistical language learning. In particular, Casillas (2008) argued that SL was insufficient because infants must have some way of knowing what chunks to form, and SL does not provide this. Gambell and Yang (2003, 2005) argued that child-directed speech does not naturally lend itself to wordlike segmentation. However, TRACX shows that unsupervised SL is sufficient to do word extraction without requiring any prior information as to what must be learned, even from a natural infant-directed speech corpus. Of course, this is not to deny that other cues enrich the infant's learning environment and, therefore, im-

prove their learning efficacy. In fact, this is most likely to be the case. What TRACX shows is that SL is minimally sufficient to support lexical segmentation for both small- and large-scale corpora.

The Broader Importance of ICR

The chunking mechanisms implemented in TRACX raise a number of issues about the nature of chunks themselves. In the traditional artificial intelligence (AI) sense, chunking means explicitly creating a new symbol that stands for a string of subsymbols, that is potentially consciously accessible, and that can be manipulated. If on time step *t*, TRACX's input matches its output (i.e., it recognizes that it has encountered that input before), then it inserts the current hidden-unit representation into half of its input on time step $t + 1$. While the hidden-unit representation of the input on time step *t* can reasonably be considered to be an internal representation of a chunk consisting of the items in the LH and RH sides of the input, this internal representation is in no way explicit, in the sense that the network could subsequently explicitly generate the items from which it was built. This contrasts with PARSER (Perruchet & Vintner, 1998), in which chunks are stored explicitly in WM and could be enumerated by the program, if necessary. This is not the case in TRACX, in which all chunks are stored in a distributed manner across the synaptic weights of the network. TRACX can recognize whether it has encountered a particular chunk of items before, but it cannot generate the chunks it has encountered at will.[10]

Conceptualizing chunks as patterns of activation across distributed hidden-unit representations also naturally implements a similarity gradient across chunks. For example, because of their close phonological similarity, which would be reflected in distributed input coding, the chunks *gaboti* and *kapodi* would generate responses that were more closely aligned than, say, *gaboti* and *pudosa*, even though none of the three chunks share any common syllables. This would be particularly hard to implement in PARSER. For the same reason, chunk interference falls naturally out of TRACX's mechanisms, whereas it has to be built in explicitly in models like PARSER.

That explicit manipulable chunks, such as *breakfast*, *cupboard*, or *doughnut*, can exist is undeniable. However, a broad interpretation of the results from TRACX would suggest that explicit chunking might not be necessary to areas of cognition in which unsupervised learning takes place. The essence of TRACX is that it is gating information flow in the network as a function of the familiarity (recognizability) of the current input.

Extraction and Transfer of Abstract Structure

We have demonstrated that TRACX can achieve sequence segmentation and chunk formation via recognition memory. One important direction for future research with this model is to explore its ability to extract abstract structure from the sequences it processes. For example, there are experiments dealing with the learning of phonological sequences that study how phonotactic-like constraints are learned by infants (e.g., Chambers, Onishi, & Fisher, 2003) and adults (Onishi, Chambers, & Fisher, 2002). These studies demonstrate the implicit learning of patterns that generalize to novel test items. The apparent abstractness of the regularities acquired implicitly can be quite striking (Chambers, Onishi, & Fisher, 2010), although it also worth noting that some authors (e.g., Perruchet & Pacteau, 1990; Pacton, Perruchet, Fayol, & Cleeremans, 2001) have argued forcefully that the appearance of abstract structures can arise from pairwise bigram associations and are, therefore, not diagnostic of intrinsic abstract structures. Similarly, there are studies that deal with the acquisition and transfer of syntactic patterns (e .g., Marcus et al., 1999) and traditional sequential RT tasks in which generalization is tested (e.g., Gupta & Cohen, 2002).

We have shown in Simulations 9 and 10 that TRACX is capable of developing internal representations whose organization reflects the overall organization of the data that it is processing. In other words, it is sensitive to simple structure in its input. In a different context, that of active-to-passive syntax transformations, Chalmers (1990) showed that the RAAM architecture is able to implicitly extract abstract structure from the input that it is processing and apply that structure to novel input. Since TRACX is a type of RAAM architecture and since, in addition, it has been shown to be able to extract elementary organizational structure from a bilingual microlanguage environment, it is reasonable to suppose that it might be able to extract more complex structure from the sequences that are presented to it. Blank et al. (1992) also show how the internal representations of a RAAM model reflect the structure in its input. Furthermore, Sirois et al. (2000) showed how simple autoassociators could transfer statistical structures extracted from sequences typical of the ones used to test infants to new test items with different surface features. Finally, other (SRN-type) connectionist architectures have also been shown to be able to transfer structures to novel sequences (e.g., Altmann, 2002; Seidenberg & Elman, 1999). This is clearly an area for future research for the TRACX model.

Modifications of TRACX

The TRACX model is a simple instantiation of the principle of ICR. It is intended primarily as a proof-of-concept model rather than a full-featured model of sequence segmentation and chunk extraction. In the simple form in which it has been presented in this article, it is able to simulate empirical results from a wide range of experiments in adult IL and infant SL. Further, a simulation of the acquisition of two microlanguages was designed to demonstrate its ability to generalize to new input and to develop internal clusters of representations that reflect category divisions in the environment.

[10] Because TRACX is a connectionist model, it does not store chunks explicitly. They are stored in a distributed manner across the weights of the network. For this reason, unlike a model that explicitly stores the chunks that it has encountered (e.g., PARSER), TRACX can recognize that a particular chunk, say, *abc*, has been encountered before, but it cannot, in its current instantiation, *generate* the chunks it has encountered. There are ways that this might be achieved, for example, by coupling a Helmholtz machine (Dayan, Hinton, Neal, & Zemel, 1995) or a deep belief network (Hinton, Osindero, & Teh, 2006) to TRACX during learning. This, however, is beyond the scope of this article. We focus on the recognition of words in the input stream and not on the network's ability to generate these words.

However, there are a number of things that the present version of TRACX cannot do. First, it does not separate out words consisting of a single syllable. There are numerous ways that this problem could be dealt with, but no such mechanisms were added to the current version of TRACX because it was felt that this would distract readers from the main point of the present article—namely, a novel mechanism of multi-item chunk extraction. This issue also relates to the issue of the size of the input processing window. At the moment, TRACX processes two input elements at a time. Ultimately, this is likely to be insufficient. The number of elements in the temporal window might, for example, more reasonably correspond to the number of items that can be stored in short-term memory.

There is also the issue of long-distance dependencies (i.e., recognizing a category of words *a–X–b*, in which *X* can be one of a number of different syllables). Cleeremans and Dienes (2008) claim no chunking models capture dependencies of this type. Indeed, in its current form, TRACX could not learn these long-distance dependencies. Again, one can imagine a modified architecture in which this would be possible. For example, there is no a priori reason to limit the number of item vectors on input to two. In any event, aside from the fact that people also find long-distance dependencies very hard to learn in impoverished semantic-free sequences typical of the studies modeled above (Mathews et al., 1989), in the present model, no attempt was made to model this type of dependency.

At least two further changes would make the model more complex but more realistic. The first involves the hard cutoff for the error criterion. As it stands, if the error measure is below a preset criterion (in this case, .4), on the next time step, the hidden-unit activations are put in the LH-side of the input to the network. This should instead be a stochastic decision based on the amount of error, specifically, the lower the error on output, the higher the probability of putting the hidden-unit activations in the LH-side of the input on the next time step.

The second change involves what can be put into the RH-side of the input to the network. Currently, the RH-side input units can only contain the representation of an individual item in the input stream. This constraint ultimately needs to be removed, so that chunked representations of several items would also be allowed in the RH-side input. So, for example, if the network has already created internal chunks for the words *under* and *cover*, it currently cannot simply put the internal representation for *under* in the LH-side input units and *cover* in the RH-side input units and directly learn the new chunk *undercover*. For the moment, it can only put the internal representation for *under* in the LH-side input units and gradually, syllable by syllable, build up, first, *under-co* and then *under-co-ver*.

One way to achieve this would be to have an identical copy of the TRACX network, but one in which no learning occurred, preprocess the sequence, looking ahead in the input sequence and replacing previously encountered chunks of items by their internal representations. Since these internal representations would have the same length as the primitive items in the sequence, they could be put into the RH-side inputs. This would also solve the problem, encountered in Simulation 5 of Giroux and Rey (2009), in which the initial chunk *kl* is never forgotten by the system in the same way that *lm* and *mn* are. For the moment, however, we have not implemented such a compound system, and consequently, the RH-side input always contains a single element from the raw input sequence.

Conclusion

TRACX is a computational model that is intended to provide a general connectionist framework for sequence processing and chunk extraction. The model implements a mechanism, ICR, which automatically extracts chunks from continuous input and uses these chunks for further processing. TRACX (a) does not have heavy memory or processing requirements, (b) relies on the recognition of previously encountered items rather than on the prediction of upcoming items, (c) is not dependent on input encoding and thus can be considered domain independent, (d) can scale from laboratory tasks to real-world situations, (e) develops internal representations that reflect structure in the world, and (f) can do simple generalization to new input.

The results produced by TRACX suggest a new approach to modeling sequence processing and chunk extraction. TRACX was successfully tested on a wide range of empirical data in the areas of adult IL and infant SL. It provides a novel, conceptually parsimonious, yet powerful, framework in which to model sequence processing and chunk extraction in general.

References

Altmann, G. T. M. (2002). Learning and development in neural networks: The importance of prior experience. *Cognition, 85,* B43–B50. doi: 10.1016/S0010-0277(02)00106-3

Anderson, B. (1999). Kohonen neural networks and language. *Brain and Language, 70,* 86–94. doi:10.1006/brln.1999.2145

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451. doi:10.1037/0033-295X.84.5.413

Aslin, R. N., Saffran, J. R., & 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

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Mahwah, NJ: Erlbaum.

Bernstein-Ratner, N. (1987). The phonology of parent–child speech. In K. E. Nelson & A. van Kleek (Eds.), *Children's language* (Vol. 6, pp. 159–174). Hillsdale, NJ: Erlbaum.

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology Section A, 36,* 209–231.

Blank, D. S., Meeden, L. A., & Marshall, J. B. (1992). Exploring the symbolic/subsymbolic continuum: A Case Study of RAAM. In J. Dinsmore (Ed.), *Closing the gap: Symbolism versus connectionism* (pp. 113–148). Mahwah, NJ: Erlbaum.

Bloom, P. (1997). Intentionality and word learning *Trends in Cognitive Sciences, 11,* 9–12.

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science: A Multidisciplinary Journal, 27,* 807–842. doi: 10.1207/s15516709cog2706_1

Brent, M. R. (1999). An efficient probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34,* 71–105. doi: 10.1023/A:1007541817488

Brent, M. R., & Cartwright, T. (1996). Distributional regularity and pho-

notactic constraints are useful for segmentation. *Cognition, 61,* 93–125. doi:10.1016/S0010-0277(96)00719-6

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Lexical segmentation: The role of sequential statistics in supervised and unsupervised models. In A. Ram & K. Eiselt (Eds.), *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 136–141). Hillsdale, NJ: Erlbaum.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33,* 111–153. doi:10.1006/cogp.1997.0649

Casillas, G. (2008). The insufficiency of three types of learning to explain language acquisition. *Lingua, 118,* 636–641. doi:10.1016/j.lingua.2007.03.007

Chalmers, D. (1990). Syntactic transformations on distributed representations. *Connection Science, 2,* 53–62.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition, 87,* B69–B77. doi:10.1016/s0010-0277(02)00233-0

Chambers, K. E., Onishi, K. H., & Fisher, C. (2010). A vowel is a vowel: Generalizing newly learned phonotactic constraints to new contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 821–828. doi:10.1037/a0018991

Christiansen, M. H. (1999). The power of statistical learning: No need for algebraic rules. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 114–119). Hillsdale, NJ: Erlbaum.

Christiansen, M. H., Allen, J., & Seidenberg, M. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13,* 221–268. doi:10.1080/016909698386528

Christiansen, M. H., Onnis, L., & Hockema, S. A. (2009). The secret is in the sound: From unsegmented speech to lexical categories. *Developmental Science, 12,* 388–395.

Christoff, K., Ream, J. M., Geddes, L. P. T., & Gabrieli, J. D. E. (2003). Evaluating self-generated information: Anterior prefrontal contributions to human cognition. *Behavioral Neuroscience, 117,* 1161–1168. doi:10.1037/0735-7044.117.6.1161

Clark, R. E., & Squire, L. R. (1998, April). Classical conditioning and brain systems: The role of awareness. *Science, 280,* 77–81. doi:10.1126/science.280.5360.77

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

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 396–421). Cambridge, England: Cambridge University Press.

Cleeremans, A. & McClelland, J. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 20,* 235–253. doi:10.1037/0096-3445.120.3.235

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163). Mahwah, NJ: Erlbaum.

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

Cottrell, G. W., & Metcalfe, J. (1991). EMPATH: Face, gender, and emotion recognition using holons. In D. Touretzky (Ed.), *Advances in neural information processing systems* (pp. 564–571). San Mateo, CA: Kaufmann.

Cottrell, G. W., Munro, P., & Zipser, D. (1988). Image compression by back-propagation: An example of extensional programming. In N. E. Sharkey (Ed.), *Advances in cognitive science* (Vol. 3, pp. 208–241). Norword, NJ: Ablex.

Crick, F. (1989). The recent excitement about neural networks. *Nature, 337,* 129–132. doi:10.1038/337129a0

Cunillera, T., Camara, E., Laine, M., & Rodriguez-Fornells, A. (2010). Words as anchors: Known words facilitate statistical learning. *Experimental Psychology, 57,* 134–141. doi:10.1027/1618-3169/a000017

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General, 128,* 165–185. doi:10.1037/0096-3445.128.2.165

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, K. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience, 14,* 1158–1173. doi:10.1162/089892902760807177

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation, 7,* 889–904. doi:10.1162/neco.1995.7.5.889

de Marcken, C. (1995). *Unsupervised acquisition of a lexicon from continuous speech* (Technical Report AI Memo No. 1558). Cambridge, MA: Massachusetts Institute of Technology.

Destrebecqz, A., & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review, 8,* 343–350. doi:10.3758/BF03196171

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science: A Multidisciplinary Journal, 16,* 41–79.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22,* 735–808. doi:10.1017/S0140525X99002186

Dominey, P. F. (1998). A shared system for learning serial and temporal structure of sensori-motor sequences? Evidence from simulation and human experiment. *Cognitive Brain Research, 6,* 163–172. doi:10.1016/S0926-6410(97)00029-3

Dominey, P., & Ramus, F. (2000). Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes, 15,* 87–127. doi:10.1080/016909600386129

Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General, 113,* 541–555. doi:10.1037/0096-3445.113.4.541

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In H. T. Bunell & W. Isardi (Eds.), *Proceedings of the Fourth International Conference on Spoken Language Processing* (pp. 1393–1396). Wilmington, DE: Alfred I. DuPont Institute.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science: A Multidisciplinary Journal, 14,* 179–211. doi:10.1207/s15516709cog1402_1

Fahlman, S. E. (1988). Faster-learning variations on back-propagation: An empirical study. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the Connectionist Models Summer School* (pp. 38–51). Los Altos, CA: Kaufmann.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9,* 59–79. doi:10.3758/BF03196257

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States, 99,* 15822–15826.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117,* 107–125. doi:10.1016/j.cognition.2010.07.005

French, R. M. (1998). A simple recurrent network model of bilingual memory. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Cognitive Science Society Conference.* Mahwah, NJ: Erlbaum, 368–737.

French, R. M., & Cleeremans, A. (2002). *Implicit learning and consciousness: An empirical, philosophical, and computational consensus in the making.* London, England: Psychology Press.

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

French, R. M., Mermillod, M., Quinn, P., & Mareschal, D. (2001). Reversing category exclusivities in infant perceptual categorization: Simulations and data. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 23nd Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum, 307–312.

Gambell, T., & Yang, C.-D. (2003). Scope and limits of statistical learning in word segmentation. In K. Moulton & M. Wolf (Eds.), *Proceedings of the 34th Northeastern Linguistic Society Meeting* (pp. 368–737). New York, NY: Stony Brook University Press.

Gambell, T., & Yang, C.-D. (2005). *Mechanisms and constraints in word segmentation.* Unpublished manuscript, Department of Linguistics, Yale University, New Haven, CT.

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., & Fried, I. (2008, October). Internally generated reactivation of single neurons in human hippocampus during free recall. *Science, 322,* 96–101. doi:10.1126/science.1164685

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science: A Multidisciplinary Journal, 33,* 260–272.

Gluck, M. A., & Granger, R. (1993). Computational models of the neural bases of learning and memory. *Annual Review of Neuroscience, 16,* 667–706. doi:10.1146/annurev.ne.16.030193.003315

Gluck, M. A., & Meyers, C. E. (1997). Psychobiological models of hippocampal function in learning and memory. *Annual Review of Psychology, 48,* 481–514. doi:10.1146/annurev.psych.48.1.481

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In D.-S. Huang, D. C. Wunsch, II, D. S. Levine, & K.-H. Jo (Eds.), *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 673–680). Berlin, Germany: Springer-Verlag. doi:10.1016/j.cognition.2009.03.008

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112,* 21–54.

Graf-Estes, K., Evans, J. L., Alibali, M., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18,* 254–260. doi:10.1111/j.1467-9280.2007.01885.x

Gupta, P., & Cohen, N. J. (2002). Theoretical and computational analysis of skill learning, repetition priming, and procedural memory. *Psychological Review, 109,* 401–448. doi:10.1037/0033-295X.109.2.401

Harris, Z. S. (1954). Distributional structure. *Word, 10,* 146–162.

Harris, Z. S. (1955). From phoneme to morpheme. *Language, 31,* 190–222. doi:10.2307/411036

Hinton, G. E. Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18,* 1527–1554. doi:10.1162/neco.2006.18.7.1527

Holmes, P., & Calmels, C. (2008). A neuroscientific review of imagery and observation in sport. *Journal of Motor Behavior, 40,* 433–445. doi:10.3200/JMBR.40.5.433-445

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia, 33,* 1419–1432. doi:10.1016/0028-3932(95)00073-C

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44,* 548–567. doi:10.1006/jmla.2000.2755

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88,* 67–85. doi:10.1037/0033-295X.88.1.67

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29,* 1–23. doi:10.1006/cogp.1995.1010

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics, 61,* 1465–1476. doi:10.3758/BF03213111

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39,* 159–207. doi:10.1006/cogp.1999.0716

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83,* B35–B42. doi:10.1016/S0010-0277(02)00004-5

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustic Society of America, 59,* 1208–1221. doi:10.1121/1.380986

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language, 5,* 44–45.

Lebiere, C., Wallach, D., & Taatgen, N. (1998). Implicit and explicit learning in ACT–R. In F. E. Ritter & R. M. Young (Eds.), *Proceedings of the Second European Conference on Cognitive Modeling* (pp. 183–189). Nottingham, England: Nottingham University Press.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology, 46,* 1–27. doi:10.1146/annurev.ps.46.020195.000245

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12,* 271–295. doi:10.1017/S0305000900006449

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999, January). Rule learning by seven-month-old infants. *Science, 283,* 77–80. doi:10.1126/science.283.5398.77

Mareschal, D., & French, R. (2000). Mechanisms of categorization in infancy. *Infancy, 1,* 59–76. doi:10.1207/S15327078IN0101_06

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

Mareschal, D., & Westermann, G. (2010). Mixing the old with the new and the new with the old: Combining prior and current knowledge in conceptual change. In S. P. Johnson (Ed.), *Neoconstructivism: The new science of cognitive development* (pp. 213–239). New York, NY: Oxford University Press.

Marcovitch, S., & Lewkowicz, D. J. (2009). Sequence learning in infants: The independent contributions of conditional probability and pair frequency information. *Developmental Science, 12,* 1020–1025. doi:10.1111/j.1467-7687.2009.00838.x

Maskara, A., & Noetzel, A. (1993). Sequence recognition with recurrent neural networks. *Connection Science, 5,* 139–152. doi:10.1080/09540099308915692

Mathews, R. C., Buss, R. R., Stanley, W. R., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1083–1100. doi:10.1037/0278-7393.15.6.1083

Mattys, S. L., Jusczyk, P., Luce, P., & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465–494. doi:10.1006/cogp.1999.0721

McClelland, J. L., Botvinich, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenber, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14,* 348–356. doi:10.1016/j.tics.2010.06.002

Mirman, D., Graf-Estes, K., & Magnus, J. S. (2010). Computational modeling of statistical learning: Effects of transitional probability versus frequency and links to word learning. *Infancy, 15,* 471–486. doi:10.1111/j.1532-7078.2009.00023.x

Mirman, D., Magnus, J. S., Graf-Estes, K., & Dixon, J. A. (2008). The link

between statistical segmentation and word learning adults. *Cognition, 108,* 271–280. doi:10.1016/j.cognition.2008.02.003

Nakatani, L. H., & Shaffer, J. A. (1978). Hearing "words" without words: Prosodic cues for word perception. *Journal of the Acoustical Society of America, 63,* 234–245. doi:10.1121/1.381719

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 756–766. doi:10.1037/0096-1523.24.3.756

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. doi:10.1037/0096-1523.24.3.756

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19,* 1–32. doi:10.1016/0010-0285(87)90002-8

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition, 83,* B13–B23. doi:10.1016/S0010-0277(01)00165-2

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General, 130,* 401–426. doi:10.1037/0096-3445.130.3.401

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition, 13,* 244–247. doi:10.1016/j.cognition.2009.07.011

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80,* 674–685. doi:10.1111/j.1467-8624.2009.01290.x

Perlman, A., Pothos, E. M., Edwards, D., & Tzelgov, J. (2010). Task-relevant chunking in sequence learning. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 649–661. doi:10.3758/MC.36.7.1299

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory and Cognition, 36,* 1299–1305.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119,* 264–275. doi:10.1037/0096-3445.119.3.264

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

Perruchet, P., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science: A Multidisciplinary Journal, 34,* 255–285. doi:10.1111/j.1551-6709.2009.01074.x

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

Perruchet, P., & Vintner, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences, 25,* 297–388.

Plunkett, K., Sinha, C., Moller, M., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols: Vocabulary growth in children and a connectionist net. *Connection Science, 4,* 293–312. doi:10.1080/09540099208946620

Pollack, J. (1989). Implications of recursive distributed representations. In D. S. Touretzky (Ed.), *Advances in neural information processing systems I* (pp. 527–536). Los Gatos, CA: Morgan Kaufmann.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence, 46,* 77–105.

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin, 133,*227–244. doi:10.1016/0004-3702(90)90005-K

Quine, W. V. O. (1960). *Word and object.* Cambridge, MA: MIT Press.

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

Robinet, V., & Lemaire, B. (2009). MDL chunker: An MDL-based model of word segmentation. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 3866–2871). Mahwah, NJ: Erlbaum.

Rolls, E. T., & Treves, A. (1997). *Neural networks and brain function,* Oxford, England: Oxford University Press.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81,* 149–169. doi:10.1016/S0010-0277(01)00132-9

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

Saffran, J. R., Johnson, E., Aslin, R. N., & Newport, E. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70,* 27–52. doi:10.1016/S0010-0277(98)00075-4

Saffran, J., Newport, E., Aslin, R. N., Tunick, R., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8,* 101–105. doi:10.1111/j.1467-9280.1997.tb00690.x

Saffran, J., & Wilson, D. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4,* 273–284. doi:10.1207/S15327078IN0402_07

Seidenberg, M. S., & Elman, J. L. (1999, April). Do infants learn grammar with algebra or statistics? *Science, 284,* 433f–434f. doi:10.1126/science.284.5413.433f

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7,* 161–193. doi:10.1007/BF00114843

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 592–608. doi:10.1037/0278-7393.16.4.592

Shanks, D. R. (2005). Implicit learning. In K. Lamberts & R. Goldstone (Eds.), *Handbook of cognition* (pp. 202–220). London, England: Sage.

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology, 61,* 273–301. doi:10.1146/annurev.psych.093008.100519

Sirois, S., Buckingham, D., & Shultz, T. R. (2000). Artificial grammar learning by infants: An autoassociator perspective. *Developmental Science, 3,* 442–456. doi:10.1111/1467-7687.00138

Sirois, S., & Mareschal, D. (2004). An interacting systems model of infant habituation. *Journal of Cognitive Neuroscience, 16,* 1352–1362. Medline doi:10.1162/0898929042304778

Sun, R. (1997). Learning, action and consciousness: A hybrid approach toward modeling consciousness. *Neural Networks, 10,* 1317–1331. doi:10.1016/S0893-6080(97)00050-6

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50,* 86–132. doi:10.1016/j.cogpsych.2004.06.001

Thiessen, E. D., Hill, E., & Saffran, J. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7,* 53–71. doi:10.1207/s15327078in0701_5

Thiessen, E. D., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39,* 706–716. doi:10.1037/0012-1649.39.4.706

Thiessen, E. D., & Saffran, J. R. (2009). How the melody facilitates the message and vice versa in infant learning and memory. *Annals of the*

New York Academy of Sciences, 1169, 225–233. doi:10.1111/j.1749-6632.2009.04547.x

van de Weijer, J. (1998). *Language input for word discovery* (Doctoral dissertation, University of Nijmegen, Nijmegen, Netherlands).

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics, 27,* 351–372. doi:10.1162/089120101317066113

Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology, 43,* 217–242. doi:10.1006/cogp.2001.0764

Received October 27, 2010
Revision received June 27, 2011
Accepted June 28, 2011 ■