TRACX 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction

Robert M. French (robert.french@u-bourgogne.fr)
LEAD-CNRS UMR 5022, Université de Bourgogne
21000 Dijon, France

Garrison W. Cottrell (gary@ucsd.edu)
Computer Science and Engineering, UCSD
La Jolla, CA 92093-0404, USA

Abstract

TRACX (French, Addyman, & Mareschal, 2011) is a recursive connectionist system that implicitly extracts chunks from sequence data. It can account for experiments on infant statistical learning and adult implicit learning, as well as real-world phoneme data, and an experiment using backward transitional probabilities that simple recurrent networks cannot account for. One criticism of TRACX, however, is the implausibility in a connectionist model of if-then-else statements. In particular, one of these statements controls what data is copied from the model's internal memory into its input, based on a hard error threshold. We, therefore, developed a more biologically-plausible version of TRACX devoid of if-then-else statements, relying only on spreading activation and without any learning error threshold. This new model, TRACX 2.0, performs essentially as well as the original TRACX model and, in addition, has two fewer parameters than the original and accounts for the *graded* nature of chunks.

Keywords: chunk extraction; statistical learning; implicit learning; recursive autoassociative memory; autoassociators.

.

Introduction

No one disputes that individuals learn to extract structure from their sensory environment. There is, however, a heated debate is over just *how* this is done. In what follows we will suggest a neurobiologically plausible, memory-based model that achieves this in the auditory domain. The model provides a strong hypothesis as to how people -- infants, as well as adults -- might segment continuous syllable streams into words. The model is an improvement of a recent connectionist memory-based model of sequence segmentation and chunking, TRACX (French, Addyman, & Mareschal, 2011). The new model improves TRACX by removing a crucial if-then-else statement in the model and replaces it with a simple connectionist mechanism.

The mainstream view of how segmentation is done, one that has held sway for the nearly two decades, is based on the notion of *prediction*. This theory supposes that individuals, based on their previous experience with the world, are constantly in the process of making predictions about what is going to happen next in their environment. In so doing, they gradually learn to align their predictions with what actually happens in the world. In order to make these predictions, they must gradually learn the probabilities of successive events in the world. We learn that a flash of lightning will invariably be followed by a clap of thunder, that a "hello" will usually be reciprocated, that a phone call will sometimes be for us, but sometimes not, that the flashing light on a police car will usually be for someone else, but occasionally for us, and so on.

This is the basis of the *transitional probability* (TP) theory of sequence segmentation. The idea is simple. In the syllable stream that an infant hears, many multi-syllable words will be repeated frequently (e.g., *bay-bee*, *mah-mee*, *bah-tul*, and so on) and, as a result, the infant will become better at predicting upcoming *within-word* syllables compared to upcoming *between-word* syllables. (The syllable pair *bay-bee* will be followed by the initial syllable of many different words, whereas as *bay* will be very frequently followed by *bee*. The infant thus learns the word *bay-bee*.) Thus, low syllable-to-syllable TPs (failures to predict) indicate word boundaries. High syllable-to-syllable TPs bind syllables together into words and facilitate their learning. An obvious connectionist candidate for this kind of transitional-probability based learning is the well-known Simple Recurrent Network (SRN, Elman, 1990).

While we don't doubt that prediction is an important aspect of cognition, there are other plausible explanations as to how infants (and adults) learn to segment continuous speech streams into words. Broadly speaking, there are four classes of models used to explain sequence segmentation and word extraction. These are:

- Predictive connectionist models, most prominent among them the SRN (Elman, 1990; Cleeremans & McClelland, 1991; Servan-Schreiber, Cleeremans, & McClelland, 1991);
- Chunking connectionist models, i.e., TRACX (French, et al., 2011);
- Symbolic hybrid models, the best known of which are probably PARSER (Perruchet & Vinter, 1998, 2002) and the Competitive Chunker (Servan-Schreiberr & Anderson, 1990)
- Normative statistical models (Frank, Goldwater, Griffiths & Tenenbaum, 2010; Goldwater, Griffiths, & Johnson, 2009; Börschinger, & Johnson, 2011).

Recently, Kurumada, Meylan, and Frank (2013) ran a series

of tests on models from each of these classes and found that "computational models that implement 'chunking' are more effective than 'transition finding' models" at reproducing segmentation in a context where the frequency of words followed a Zipfian distribution (e.g., words in real natural language). TRACX was singled out as the model that best captured human word-segmentation performance in a Zipfian context.

However, even though French et al. (2011) criticize the lack of neurobiological plausibility of competing non-connectionist models of sequence segmentation, one of the key features of their own model undermines its claim to neurobiological plausibility. This feature (an if-then-else switch) plays a crucial role in ensuring that the network can re-use syllable chunks that it has detected in the input. In what follows we show that this flaw can be overcome and develop a new, simpler implementation of the original TRACX model, which we call TRACX 2.0. This modified version of TRACX not only replaces the problematic feature with a simple, neurobiologically sound mechanism, but also requires two fewer parameters than the original model. We also show that TRACX 2.0 produces qualitatively the same results as the original TRACX model on five datasets for infants and adults.

TRACX

The architecture of the TRACX model is explained in detail in French et al. (2011). Here we present a brief summary of the architecture.

TRACX is a member of the Recursive Auto-Associative Memory (RAAM) family of connectionist architectures (Pollack, 1990; Blank, Meeden & Marshall, 1992). It is a three-layer (input-hidden-output) connectionist autoassociator whose key ability is to learn to recognize when it has seen pairs of input items before.

Autoassociators gradually learn to produce output that is identical to their input. This means that items that they have seen frequently on input will be accurately reproduced on output, unlike items that have not been seen by the autoassociator before, or have only been seen infrequently. This provides the autoassociator with a simple way of determining whether or not it has previously encountered the vector of values currently on its input: if the output is very different from the input, it is novel. If it is very close, it is known. This signal is also the error signal that drives the weight changes, making the output more similar to the input.

Plausibility of Autoassociation

Autoassociators have a long history in the computational modeling of cognition. The first model to make a lasting mark was Anderson's Brain State in a Box (BSB) model (Anderson, Silverstein, Ritz and Jones, 1977). This model had no hidden layer and could not learn internal representations of its input. Ackley, Hinton, and Sejnowski (1985) were the first to add a hidden layer to their autoassociators, thereby allowing them learn compact representations of their input (hence these models are also called *autoencoders*).

Today, the psychological and biological plausibility of autoassociation is widely accepted (Rolls & Treves, 1997). Autoassociators have been successfully used as psycho-biologically plausible models in many areas of cognition. For example, Mareschal, French, & Quinn (2000) and French, Mareschal, Mermillod & Quinn (2004) developed an autoassociator model of infant categorization based on the autoassociative principles of Sokolov (1963) and others. Other psycho-biologically plausible models using autoassociators include models of face perception (Cottrell & Metcalfe, 1991), of hippocampal/episodic memory (Metcalfe, Cottrell & Mencl, 1992; Gluck & Granger, 1993), of serial recall memory (Farrell & Lewandowsky, 2002), and infant habituation (Sirois & Mareschal, 2004).

The Architecture of TRACX

The original TRACX autoassociator is constructed as follows. The input layer is divided into a Left-Hand Side (LHS) and a Right-Hand Side (RHS), each with the same number of units. Being an autoassociator, it, of course, has the same number of inputs and outputs; being a RAAM, the hidden layer has half as many units as the input layer, which allows the hidden layer to be copied back to the input layer and combined with the next input. Aside from the potential copy-back, the network is fully feedforward. The weights are changed by standard backpropagation based on the error between the input and output. Learning stops when the network is below an error threshold of 0.4. The input data was encoded with bipolar units (-1 and 1).

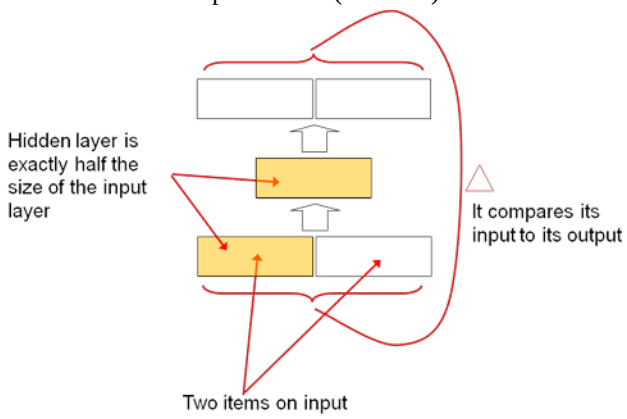

Figure 1. TRACX architecture: a 2N-N-2N autoassociator.

How TRACX Works

The easiest way to explain the architecture of TRACX is by means of an example. Assume we have a language that consists of four 3-syllable words made of distinct syllables: *abc*, *def*, *ghi*, and *jkl*. We then present to the network a continuous syllable stream made up of these words:

abcjkldefghidefabcdefabcghiabcdefabc...

These syllables are read by TRACX in sequential order. So, (the bipolar encoding of) *a* is put into the LHS and *b* into

the RHS. From the input layer, activation spreads forward to the output layer. The difference between the input and the output determines the error signal, and the weights of the system are modified by backpropagation based on this error. Initially, this error will be high. If the error is above threshold (in this case, 0.4), the value in the RHS will be shifted into the LHS and the next syllable -- in this case, *c* -- will be shifted into the RHS. If, on the other hand, the network has seen the *a-b* input many times, its output will be close to *a-b*, and the error will be small. The fact that *a*

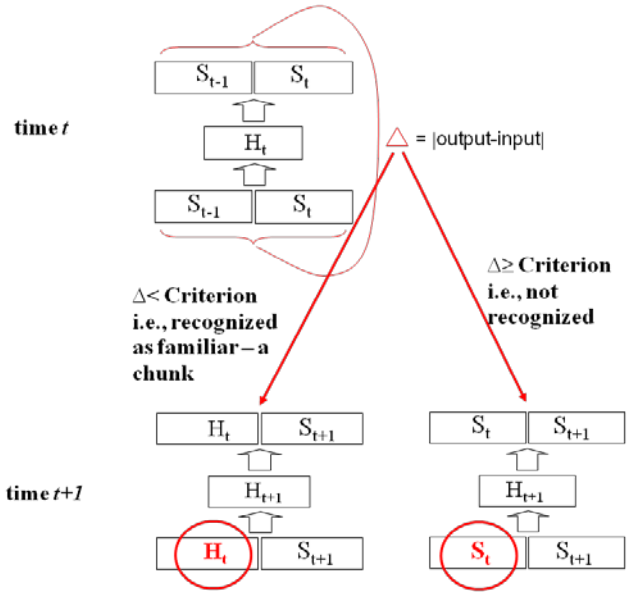

Figure 2. TRACX behavior at *t*+1 depends on error at *t*.

and *b* occur together many times in the input is another way of saying that *a-b* form a *chunk*. Once the network "recognizes" that *a-b* is a chunk (because the autoassociative error for *a-b* is below the error threshold), its behavior changes. On the following time step, instead of putting *b* into the LHS, it puts *the hidden unit representation of a-b* (call it H_{ab}) into the LHS. As before, it puts the next syllable, *c*, into the RHS (Figure 2). Now it attempts to autoassociate H_{ab}-*c*. Since, in fact, it will see *a-b-c* many times, it will eventually learn to autoassociate (and chunk) H_{ab}-*c*. It will not chunk further because *c* can be followed by any one of four different syllables, so the error will always be high.

This switch in behavior based on the error threshold is the if-then-else we would like to eliminate in our new version of the model.

Testing TRACX: words vs. partwords

Words are syllable groups that are "bound together" as a chunk. On the other hand, *partwords* are typically made up of the final syllable of one word and the initial syllable(s) of another word. It turns out that humans - infants and adults - learn to segment words from a continuous syllable stream better than partwords. In the example in the previous section, *abc, def, ghi, and jkl* are words, whereas *cjk, lde, fgh, ide, cde, cgh*, etc. are partwords. In all five of the experiments we model below, three involving infants and two involving adults, words are learned significantly better than partwords.

After the model has been trained, it is tested on a stream of data similar to the one it was trained on (as are humans and babies). There are numerous ways in which the output error of words and partwords could have been calculated. The one we chose is as follows. An item (in this case, a 3-syllable word or partword), say *abc* (or partword, *cde*), is given to the network. *a* and *b* are put on the input. This input is fed through to the hidden layer, which produces a vector of hidden-unit activations. For words/partwords longer than two syllables, as is the case here, this hidden-unit vector is then put in the LHS of the network and *c* is put in the RHS. This is then fed through to the output, and the maximum of the absolute values of the error across all output nodes is the error measure for item *abc*.

Improving TRACX

The problem with the original TRACX model is that if-then-else statements are not palatable in a connectionist context where it is unclear how such a branching behavior can be implemented (but such behavior can be learned - see Cottrell & Tsung, 1993).

The operation of the original TRACX model cannot function without two conditionals in its midst. The first is by far the most important:

IF Network error is greater than the Error Threshold,
THEN Put the element in the RHS in the LHS
ELSE Put the hidden unit vector into the LHS.

The network then grabs the next element in the sequence and puts it into the RHS and feeds what is on the input layer through to the output layer of the network. At this point the second conditional is applied:

IF LHS contains the previous Hidden-unit vector
THEN Do a backpropagation pass 25% of the time
ELSE Do a backpropagation pass.

This was a simple means of ensuring that the network, like people, place less emphasis on internally generated input compared to input from the sensory interface. In TRACX 2.0 neither of these statements is necessary.

TRACX without Tears

The improved version of TRACX is based on a simple observation concerning the graded nature of chunking. In the previous version of TRACX, the two elements on the input were either considered by the network to be chunked (if the error on output was below the Error Threshold) or they were not. There was no middle ground. But this is cognitively unrealistic, since, in fact, chunks are *graded*. By this we mean that some chunks are more "chunked" than others. To illustrate this, consider some chunks made, not from syllables, but from words. For example, almost no one hears the component words "cup" and "board" in the word "cupboard". The word has been with us since the late 14th century when it meant a board on which cups and other similar objects were placed. But over the course of 500

years, the two words, "cup" and "board", have fused so completely that we no longer hear them as separate entities. On the other hand, newer compound words, such as "smartphone", "mousepad", or "congresswoman" are at the other extreme: these words are weakly chunked; we still clearly hear their component words. These words are far less strongly chunked than words like "breakfast", "football" or "cupboard".

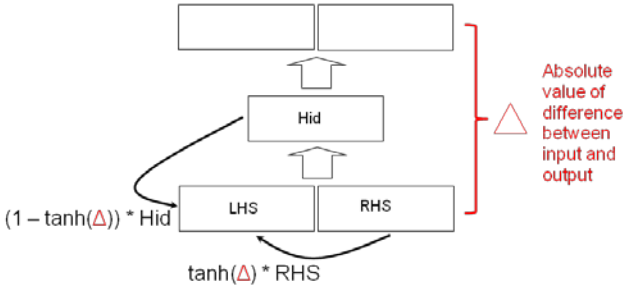

Figure 3. Information transfer in TRACX 2.0.

Thus, beyond the neurobiological implausibility of the if-then-else statement in the original TRACX, the dichotomous nature of chunks is dubious. We have changed this in TRACX 2.0. In the new model the contents of the LHS at time *t*+1 is a weighted sum of the hidden-layer vector and the RHS:

$$LHS = \left(1 - \tanh(\Delta)\right) * Hid + \tanh(\Delta) * RHS$$

where Δ is the absolute value of the maximum (component-wise) error on the output (hence it ranges from 0 to 2, and, therefore, tanh(Δ) ranges between 0 to 1) at time *t*. If Δ is small ("I've seen these two items together in the input numerous times before"), then the contribution from the hidden layer will be large. If Δ is large, most of the contribution to the LHS will come almost exclusively from the RHS (Figure 3). This can easily be implemented via a unit that takes Δ as input, and then multiplicatively gates the connections to the LHS (positively with the RHS, and negatively with a bias of 1 on the hidden layer).

This weighted sum of activation sent to the LHS removes the problematic if-then-else statements in the original TRACX, and implements the graded notion of chunks. In addition, we have found that modifying the amount of backpropagation, as in the original TRACX model is unnecessary. Chunks become stronger over time the more they are encountered, as we know occurs in humans (perhaps our children do not hear "smart" and "phone" when they refer to their "smartphone", but those of us of a certain age still do).

Testing TRACX 2.0

We tested TRACX 2.0 on five of the data sets on which the original TRACX model was tested (see French et al., 2011). In what follows we will briefly consider each of these data sets and discuss the performance of TRACX, TRACX 2.0, and an SRN on this data. These experiments are: Saffran et al. (1996), Aslin et al. (1998), Perruchet & Desaulty (2008), forward TPs, Perruchet & Desaulty (2008), backward TPs, and French et al. (2008), Equal TPs. In all of these experiments with human participants, words were learned better than partwords. This is also the case for both TRACX and TRACX 2.0, but not the case for the SRN.

Saffran, Aslin & Newport (1996) This is the seminal paper in infant syllable-sequence segmentation. Six different words were used, each with 3 distinct syllables from a 12-syllable alphabet. A random sequence of 90 of these words (270 syllables) with no immediate repeats or pauses between words was presented twice to 8-month-old infants. After this familiarization period, the infants heard a word from the familiarization sequence and a partword from that sequence. A head-turn preference procedure was used to show that infants had a novelty preference for partwords. The conclusion of the authors was that the infants had learned words better than partwords.

We simulated this experiment with TRACX, TRACX 2.0 and an SRN using the same number of words drawn from a 12-syllable alphabet. The familiarization sequence was the same length as the one the infants heard. All three models learned words better than partwords, although the SRN is considerably farther from human performance than TRACX or TRACX 2.0 (Table 1).

Aslin, Saffran & Newport (1998) In Saffran et al. (1996) there was a confound -- namely, words were heard three times as often as partwords. Aslin et al. designed an experiment that was meant to remove the unbalanced frequency of words and partwords. There were four 3-syllable words, two of which occurred twice as often in the familiarization sequence as the other two. Thus, the partwords spanning the two high-frequency words would have the same overall frequency in the familiarization sequence as the low-frequency words. The same head-turn preference procedure showed, again, that infants had a novelty preference for partwords. The conclusion of the authors was that the infants had learned words better than partwords.

Once again, we designed a set of words exactly like those used in Aslin et al. The length of the familiarization sequence was also identical to that used in Aslin et al. We tested TRACX, TRACX 2.0 and an SRN on words and partwords from this sequence, and found that all three networks learned words better than partwords, although the SRN is, again, farther from human performance than either TRACX or TRACX 2.0 (Table 1).

Perruchet & Desaulty (2008), forward TPs This is an experiment on adults. Nine 2-syllable words were constructed from 12 syllables. The familiarization string was 1035 words long, and each word occurred 115 times. The internal forward transitional probability between syllables in each word was 1. Not surprisingly, participants learned words better than partwords. We simulated this experiment by using 2-syllable words drawn from a 12-syllable alphabet to construct a familiarization sequence identical in length to the one used by Perruchet & Desaulty.

TRACX, TRACX 2.0 and the SRN learned words better than partwords. Again, the performance of the SRN was the farthest from human data (Table 1).

Perruchet & Desaulty (2008), Backward TPs This experiment, run on adults, is of crucial importance. Perruchet & Desaulty were the first to realize that *backward* TPs could serve as a segmentation cue. To illustrate the contrast between backward and forward TPs, consider the bigram *qu* in English. Given a *q*, the probability that it will be followed by a *u* is, essentially, 1. However, given a *u*, the probability that it will be *preceded* by a *q* (i.e., the backward TP) is only 0.01. Backward TPs can, in some cases, actually be higher than forward TPs. Consider the extremely common suffix *ez* in French, as in, "Parl*ez*-vous français?" The probability that, given a *z*, it will be preceded by an *e* is approximately 0.84, whereas the probability that an "e" will be followed by a "z" is a mere 0.027 (http://www.lexique.org/listes/liste_bigrammes). Perruchet & Desaulty created a set of 2-syllable words that made up a familiarization sequence in which the first syllable of each word was perfectly predicted by second syllable (i.e., backward TP = 1), whereas the second syllable was only very weakly predicted by the first syllable. They showed that under these conditions, participants still recognized words better than partwords.

We encoded the Perruchet & Desaulty vocabulary and generated sequences identical to theirs in which the word chunking cues were exclusively the backward TPs between the two syllables of the words. Since SRNs are sensitive only to forward prediction, we predicted that the SRN would fail on this data set. This proved to be the case. For this data the SRN learned partwords significantly better than words. On the other hand, both TRACX and TRACX 2.0, once again, recognized words better than partwords (Table 1). The reason for this is clear. Both of these models rely on the recognition of previously seen chunks. They are not concerned with TPs, whether forward or backward, between the syllables comprising a word. They rely on remembering having seen the pairs of syllables making up words, something that does not require TPs.

French, Addyman & Mareschal, (2011), Equal forward TPs In this experiment, run on infants, all forward TPs between syllables and between the words in the language were identical. Backward TPs within words were 1 and backward TPs between words were 0.25. Each word is associated with two partwords. French et al. determined by means of a head-turn preference procedure identical to the one used by Aslin et al. (1998), that infants exposed to this "language" learn words significantly better than partwords.

	Segmentation cues	Score type	Words learned significantly better than Partwords? (proportion better)			
			Humans	TRACX	TRACX 2.0	SRN
Saffran et al. (1996)	Freq. + Fwd TPs	looking time	Yes 0.06	Yes 0.08	Yes 0.11	Yes 0.68
Aslin et al. (1998)	Fwd TPs	looking time	Yes 0.04	Yes 0.13	Yes 0.09	Yes 0.58
Perruchet & Desaulty (2008). Expt. 2	Fwd TPs	% correct responses	Yes 0.34	Yes 0.38	Yes 0.17	Yes 0.80
Perruchet & Desaulty (2008). Expt. 2	Bkwd TPs	% correct responses	Yes 0.22	Yes 0.32	Yes 0.05	No -0.10
Equal TP	Freq. + Bkwd TPs	looking time	Yes 0.13	Yes 0.50	Yes 0.06	Yes 0.05

Table 1. Proportion of words learned better than partwords for the three models and humans on five experimental data sets.

We expected that the SRN would also learn words better than partwords, because of the greater frequency of words. However, in the absence of forward TP information, we also expected it to perform far less well than it did with the Saffran et al. (1996), the Aslin et al. (1998) and Perruchet & Desaulty (2008), forward-TP data sets. This is, indeed, what we observed (Table 1). Both TRACX and TRACX 2.0 also learn words better than partwords. However, the performance of original TRACX model is quite far from human performance, unlike TRACX 2.0, which is much closer to human performance on this data set (Table 1).

In Table 1, we use a "proportion better" measure to compare model results and empirical data. This is a relative-difference measure that can be applied equally well to error measures, to looking times, or to proportion-correct scores. (See French et al., 2011, footnote 5, p. 422, for a detailed justification of this measure.) It is calculated by taking difference of the measures for partwords and words and dividing this difference by the sum of these two measures.

Finally, we compared the human data from the five experiments and the average overall performance of TRACX, TRACX 2.0,and the SRN. The performance of TRACX and TRACX2.0 are essentially equivalent over the set of problems (Figure 4). Even though the performance of TRACX is closer to human data on three of the problems in Table 1, TRACX 2.0 is better on two of two others, and over all five experiments, the performance of the two models is similar. By contrast, the SRN's performance is considerably farther from human data on these 5 tasks.

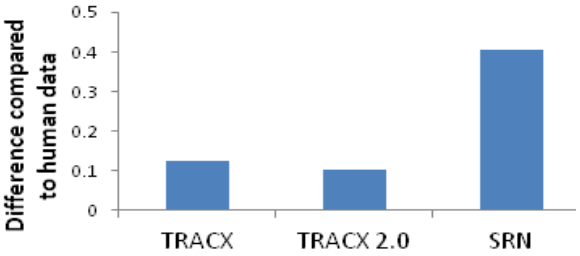

Figure 4. The three models' match to human performance, averaged over 5 experiments.

Conclusion

This article is not claiming that TRACX 2.0's performance is superior to that of the original TRACX. Rather, it is sufficient for our purposes that TRACX 2.0 performs in a qualitatively similar manner compared to the original TRACX model. What *is* important is that TRACX 2.0 no longer requires the inclusion of an error-threshold parameter, nor an external world/internal representation parameter governing how often learning takes place, and still performs in a manner qualitatively similar to TRACX. The new model, like TRACX, is a recursive autoassociator but, unlike its predecessor, it makes use only of spreading activation and error backpropagation, which has been shown to be isomorphic to the neurobiologically plausible mechanism of contrastive Hebbian learning (O'Reilly & Munakata, 2000). This is a considerable improvement over the original model for a number of reasons, namely: i) it considerably increases the neurobiological plausibility of the model; ii) it treats chunks in an appropriate, graded manner, rather than dichotomously; iii) it reduces the number of free parameters in the model by two; and, finally, iv) these changes do not significantly degrade the original model's performance.

Acknowledgments

This work was financed in part by a grant (ANR-10-0056) from the French National Research Agency to the first author, and NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center, to the second author.

References

Ackley, D.H., Hinton, G.E., and Sejnowski, T.J. (1985) A learning algorithm for Boltzmann machines. *Cog. Science* **9**:147-169.

Anderson, J.A., Silverstein, J.W., Ritz, S.A. and Jones, R.S. (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, **84**(5):413-451.

Blank, D.S., Meeden, L.A., and Marshall, J.B. (1992). Exploring the Symbolic/Subsymbolic continuum: A Case Study of RAAM. In J. Dinsmore (ed) *Closing the Gap: Symbolism vs. Connectionism.* Mahwah: NJ: LEA, pp. 113-148.

Börschinger, B., & Johnson, M. (2011). A particle filter algorithm for Bayesian word segmentation. Proceedings of the Australasian Language Technology Association, pp. 10–18.

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

Cottrell, G.W., Metcalfe, J. (1991) EMPATH: face, gender and emotion recognition using holons. D. Touretzky (Ed.) *Advances in Neural Information Processing Systems 3* (pp. 564-571), San Mateo, CA: Morgan Kaufmann.

Cottrell, G.W. and Tsung, F-S. (1993) Learning simple arithmetic procedures. *Connection Science*, **5**(1):37-58.

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

Farrell, S. & Lewandowsky, S. (2002) An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin and Review*, **9**:59-79.

Frank, M., Goldwater, S., Griffiths, T., Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition* **117(**2):107-125.

French, R. M., Addyman, C., and Mareschal, D. (2011). TRACX: A Recognition-Based Connectionist Framework for Sequence Segmentation and Chunk Extraction. *Psychological Review, 118(4)*, 614-636.

French, R. M., Mareschal, D., Mermillod, M. and 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.

Gluck, M.A. and Granger, R. (1993) Computational models of the neural bases of learning and memory. *Annual Review of Neuroscience*, **16**:667-706.

Goldwater, S., Griffiths, T.L. and Johnson. M. (2009) A Bayesian framework for word segmentation : Exploring the effects of context. *Cognition*, **112**(1):21-54

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition* **127**:439–453

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

Metcalfe, J., Cottrell, G.W., & Mencl, W.E. (1992). Cognitive Binding: A computational-modeling analysis of the distinction between implicit and explicit memory systems. *Journal of Cognitive Neuroscience* **4**(3):289-298.

O'Reilly, R., and Munakata, Y. (2000). *Computational explorations in cognitive neuroscience.* MIT Press.

Perruchet, P. and Desaulty, S. (2008) A role for backward transitional probabilities in word segmentation? *Memory and Cogntion*, **36**(7):1299-1305.

Perruchet, P. and Vinter, A. (2002) The Self-Organizing Consciousness. *Behavioral and Brain Sciences*, **25**:297- 330.

Pollack, J. (1990) Recursive Distributed Representations *Artificial Intelligence*, **46**:77-105.

Rolls, E.T. and Treves, A. (1997) *Neural Networks and Brain Function*, Oxford: Oxford University Press.

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

Servan-Schreiber, D. & Anderson, J.R. (1990) Learning artificial grammars with competitive chunking. *JEP:LMC*, *16*, 592–608.

Servan-Schreiber, D., Cleeremans, A. and McClelland, J.L. (1991) Graded state machines: The representation of temporal contingencies in simple recurrent networks *Machine Learning* **7**(2):161-193.

Sokolov, E. N. (1963). *Perception and the conditioned reflex.* Hillsdale, NJ: LEA.