Implicit learning of non-adjacent dependencies

A graded, associative account

Luca Onnis, Arnaud Destrebecqz, Morten H. Christiansen, Nick Chater, & Axel Cleeremans
Nanyang Technological University / Université Libre de Bruxelles / Cornell University / University of Warwick / Université Libre de Bruxelles

Language and other higher-cognitive functions require structured sequential behavior including non-adjacent relations. A fundamental question in cognitive science is what computational machinery can support both the learning and representation of such non-adjacencies, and what properties of the input facilitate such processes. Learning experiments using miniature languages with adult and infants have demonstrated the impact of high variability (Gómez, 2003) as well as nil variability (Onnis, Christiansen, Chater, & Gómez (2003; submitted) of intermediate elements on the learning of nonadjacent dependencies. Intriguingly, current associative measures cannot explain this U shape curve. In this chapter, extensive computer simulations using five different connectionist architectures reveal that Simple Recurrent Networks (SRN) best capture the behavioral data, by superimposing local and distant information over their internal 'mental' states. These results provide the first mechanistic account of implicit associative learning of non-adjacent dependencies modulated by distributional properties of the input. We conclude that implicit statistical learning might be more powerful than previously anticipated.

Most routine actions that we perform daily such as preparing to go to work, making a cup of coffee, calling up a friend, or speaking are performed without apparent effort and yet all involve very complex sequential behavior. Perhaps the most apparent example of sequential behavior – one that we tirelessly perform since we were children – involves speaking and listening to our fellow humans. Given the relative ease with which children acquire these skills, the complexity of learning sequential behavior may go unseen: At first sight, producing a sentence merely consists of establishing a chain of links between each speech motor action and the next, a simple addition of one word to the next. However, this characterization falls short of one important property of structured sequences. In language, for instance, many

DOI 10.1075/sibil.48.100nn

syntactic relations such as verb agreement hold between words that may be several words apart, such as for instance in the sentence *The dog that chased the cats is playful*, where the number of the auxiliary *is* depends on the number of the non-adjacent subject *dog*, not on the nearer noun *cats*.

The presence of these nonadjacent dependencies in sequential patterns poses a serious conundrum for learning-based theories of language acquisition and sequence processing in general. On the one hand, it appears that children must learn the relationships between words in a specific language by capitalizing on the local properties of the input. In fact, there is increasing empirical evidence that early in infanthood learners become sensitive to such local sequential patterns in the environment: For example, infants can exploit high and low transitional probabilities between adjacent syllables to individuate nonsense words in a stream of unsegmented speech (Saffran, Aslin, & Newport, 1996; Saffran, 2001; Estes, Evans, Alibali & Saffran, 2007). Under this characterization, it is possible to learn important relations in language using local information. On the other hand, given the presence of nonadjacent dependencies in language acquisition (Chomsky, 1959) as well as in sequential action (Lashley, 1951) associative mechanisms that rely exclusively on adjacent information would appear powerless. For instance, processing an English sentence in a purely local way would result in errors such as **The dog that chased the cats are playful*, because the nearest noun to the auxiliary verb *are* is the plural noun *cats*. An outstanding question for cognitive science is thus whether it is possible to learn and process serial nonadjacent structure in language and other domains via associative mechanisms alone.

In this paper, we tackle the issue of the implicit learning of linguistic nonadjacencies using a class of associative models, namely connectionist networks. Our starting point is a set of behavioral results on the learning of nonadjacent dependencies initiated by Rebecca Gómez. These results are interesting because they are both intuitively counterintuitive, and because they defy any explicit computational model to our knowledge. Gómez (2002) found that learning non-local $A_i\_B_i$ relations in sequences of spoken pseudo-words with structure *A X B* is a function of the variability of *X* intervening items: infants and adults exposed to more word types filling the *X* category detected the non-adjacent relation between specific A_i and specific B_i words better than learners exposed to a small set of possible *X* words. In follow-up studies with adult learners, Onnis, Christiansen, Chater, and Gómez (2003; submitted) and Onnis, Monaghan, Christiansen, and Chater (2004) replicated the original Gómez results, and further found that non-adjacencies are better learned when no variability of intervening words from the *X* category occurred. This particular *U shape learning curve* also holds when completely new intervening words are presented at test (e.g. A_i *Y* B_i), suggesting that learners distinguish nonadjacent relations indepen-

dently of intervening material, and can generalize their knowledge to novel sentences. In addition, the U shape was replicated using abstract visual shapes, suggesting that similar learning and processing mechanisms may be at play for *non-linguistic* material presented in a *different* sensory domain. Crucially, it has been demonstrated that implicit learning of nonadjacent dependencies is significantly correlated with both offline comprehension (Misyak & Christiansen, 2012) and online processing (Misyak, Christiansen & Tomblin, 2010a, b) of sentences in natural language containing long-distance dependencies.

The above results motivate a reconsideration of the putative mechanisms of non-adjacency learning in two specific directions: first, they suggest that non-adjacency learning may not be an all-or-none phenomenon and can be modulated by specific *distributional properties* of the input to which learners are exposed. This in turn would suggest a role for implicit associative mechanisms, variably described in the literature under terms as statistical learning, sequential learning, distributional learning, and implicit learning (Perruchet & Pacton, 2006; Frank, Goldwater, Griffiths, & Tenenbaum, 2010). Second, the behavioral U shape results would appear to challenge virtually all current associative models proposed in the literature. In this paper we thus ask whether there is at least one class of implicit associative mechanisms that can capture the behavioral U shape. This will allow us to understand in more mechanistic terms how the presence of embedded variability facilitates the learning of non-adjacencies, thus filling the current gap in our ability to understand this important phenomenon. Finally, to the extent that our computer simulations can capture the phenomenon without requiring explicit forms of learning, they also provide a proof of concept that implicit learning of non-adjacencies is possible, contributing further to the discussion of what properties of language need necessarily to be learned explicitly.

The plan of the paper is as follow: we first briefly discuss examples of nonadjacent structures in language and review the original experimental study by Gómez and colleagues, explaining why they challenge associative learning mechanisms. Subsequently we report on a series of simulations using Simple Recurrent Networks (SRNs) because they seem to capture important aspects of serial behavior in language and other domains (Botvinick & Plaut, 2004, 2006; Christiansen & Chater, 1999; Cleeremans, Servan-Schreiber, & McClelland; 1989; Elman, 1991, among others). Further on, we test the robustness of our SRN simulations in an extensive comparison of connectionist architectures and show that only the SRNs capture the human variability results closely. We discuss how this class of connectionist models are able to entertain both local and distant information in graded, superimposed representations on their hidden units, thus providing a plausible implicit associative mechanism for detecting non-adjacencies in sequential learning.

The problem of detecting nonadjacent dependencies in sequential patterns

At a general level, non-adjacent dependencies in sequences are pairs of mutually dependent elements separated by a varying number of embedded elements. We can consider three prototypical cases of non-local constraints (from Servan-Schreiber, Cleeremans, & McClelland, 1991) and we can ask how an ideal learner could correctly predict the last element (here letter) of a sequence, given knowledge of the preceding elements. Consider the three following sequences:

(1) L KPS V versus L KPS M

(2) L KPS V versus P GBP E

(3) L KPS V versus P KPS E

As for (1), it is impossible to predict V versus M correctly because the preceding material "L KPS" is exactly identical. Example (2), on the other hand is trivial, because the last letter is simply contingent on the penultimate letter ('V' is contingent on 'S' and E' is contingent on 'P'). Example (3), the type investigated in Gómez (2002), is more complex: the material 'KPS' preceding 'V' and 'S' does not provide any relevant information for disambiguating the last letter, which is contingent on the initial letter. The problem of maintaining information about the initial item *until* it becomes relevant is particularly difficult for any local prediction-driven system, when the very same predictions have to be made on each time step in either string for each embedded element, as in (3).

Gómez (2002) noted that many relevant examples of non-local dependencies of type (3) are found in natural languages: they typically involve items belonging to a relatively small set (functor words and morphemes like *am, the, -ing, -s, are*) interspersed with items belonging to a much larger set (nouns, verbs, adjectives). This asymmetry translates into sequential patterns of highly invariant non-adjacent items separated by highly variable material. For instance, the present progressive tense in English contains a discontinuous pattern of the type "tensed auxiliary verb + verb stem + -ing suffix", e.g. *am cooking*, *am working*, *am going*, etc.). This structure is also apparent in number agreement, where information about a subject noun is to be maintained active over a number of irrelevant embedded items before it actually becomes useful when processing the associated main verb. For instance, processing the sentence:

(4) *The dog that chased the cats is playful*

requires information about the singular subject noun "dog" to be maintained over the relative clause "that chased the cats", to correctly predict that the verb "is" is singular, despite the fact that the subordinate object noun immediately adjacent to it,

"cats", is plural. Such cases are problematic for associative learning mechanisms that process local transition probabilities (i.e. from one element to the next) alone, precisely because they can give rise to spurious correlations that would result in erroneously categorizing the following sentence as grammatical:

(5) **The dog that chased the cats are playful*

In other words, the embedded material appears to be wholly irrelevant to mastering the non-adjacencies: not only is there an infinite number of possible relative clauses that might separate *The dog* from *is*, but also structurally different non-adjacent dependencies might share the very same embedded material, as in (4) above versus

(6) *The dogs that chased the cats are playful*

Gómez exposed infants and adults to sentences of a miniature language intended to capture such structural properties, namely with sentences of the form $A_iX_jB_i$, instantiated in spoken nonsense words. The language contained three families of non-adjacencies, denoted $A_1\_B_1$, (*pel_rud*), $A_2\_B_2$ (*vot_jic*), and $A_3\_B_3$ (*dak_tood*). The set-size from which the embedded word X_j could be drawn was manipulated in four between-subjects conditions (set-size = 2, 6, 12, or 24; see Figure 1, columns 2–5). At test, participants had to discriminate between expressions containing correct non-adjacent dependencies, (e.g. $A_2X_1B_2$, *vot vadim jic*) from incorrect ones (e.g. $^*A_2X_1B_1$, *vot vadim rud*).

This test thus required fine discriminations to be made, because even though incorrect sentences were novel three-word sequences (or trigrams), both single-word and two-word (bigrams) sequences (namely, A_2X_1, X_1B_2, X_1B_1) had appeared in the training phase. In addition, because the same embeddings appeared in all three pairs of non-adjacencies with equal frequency, all bigrams had the same frequency within a given sets-size condition. In particular, the transitional probability of any *B* word given the middle word *X* was the same, for instance, P(*jic*|*vadim*) = P(*rud*|*vadim*)= .33, and so it was not possible to predict the correct grammatical string based on knowledge of adjacent transitional probabilities alone. Gómez hypothesized that if adjacent transitional probabilities were made weaker, the non-adjacent invariant frame $A_i\_B_i$ might stand out as invariant. This should happen when the set-size of the embeddings is larger, hence predicting better learning of the non-adjacent dependencies under conditions of high embedding variability. Her results supported this hypothesis: participants performed significantly better when the set-size of the embedding was largest, i.e. 24 items.

An initial verbal interpretation of these findings by Gómez (2002) was that learners detect the nonadjacent dependencies when they become invariant enough with respect to the varying embedded *X* words. This interpretation thus suggests that – while learners are indeed attuned to distributional properties of the local

environment – they also learn about which source of information is most likely to be useful – in this case adjacent or non-adjacent dependencies. Gómez proposed that learners may capitalize on the most statistically reliable source of information in an attempt to reduce uncertainty about the input (Gómez, 2002). In the context of sequences of items generated by artificial grammars, the cognitive system's relative sensitivity to the information contained in bigrams, trigrams or in long-distance dependencies may therefore hinge upon the statistical properties of the specific environment that is being sampled.

In follow-up studies, Onnis et al. (2003; submitted) were able to replicate Gómez' experiment with adults, and added a new condition in which there is only one middle element ($A_1X_1B_1$, $A_2X_1B_2$, $A_3X_1B_3$; see Figure 1, column 1). Under such condition, variability in the middle position is thus simply eliminated, thus making the *X* element invariable and the *A_B* non-adjacent items variable. Onnis et al. found that this flip in what changes versus what stays constant again resulted in successful learning of the non-adjacent contingencies. Interestingly, learning in Onnis et al.'s set-size 1 condition does not seem to be attributable to a different mechanism involving rote learning of whole sentences. In a control experiment, learners were required to learn not three but six nonadjacent dependencies and one *X*, thus equating the number of unique sentences to be learned to those in set-size 2, in which learning was poor. The logic behind the control was that if learners relied on memorization of whole sentences on both conditions, they should fail to learn the six nonadjacent dependencies in the control set-size 1. Instead, Onnis et al. found that learners had little problem learning the six non-adjacencies, despite the fact that the language control set-size 1 was more complex (13 different words and 6 unique dependencies to be learned) than the language of set-size 2 (7 words and three dependencies). This control thus ruled out a process of learning based on mere memorization and suggested that the invariability of *X* was responsible for the successful learning. A further experiment showed that learners endorsed the correct non-adjacencies even when presented with completely new words at test. For instance, they were able to distinguish $A_1Y_1B_1$ from $A_1Y_1B_2$, suggesting that the process of learning non-adjacencies leads to correct *generalization* to novel sentences.

In yet another experiment, they replicated the U shape and generalization findings with visually presented pseudo-shapes. Taken together, Gómez's and Onnis et al.'s results indicate that learning is best either when there are many possible intervening elements or when there is just one such element, with considerably degraded performance for conditions of intermediate variability (Figure 2). For the sake of simplicity, from here on we collectively refer to all the above results as the 'U shape results'. Before moving to our new set of connectionist simulations, the next section evaluates whether current associative measures of implicit learning can predict the U shape results.

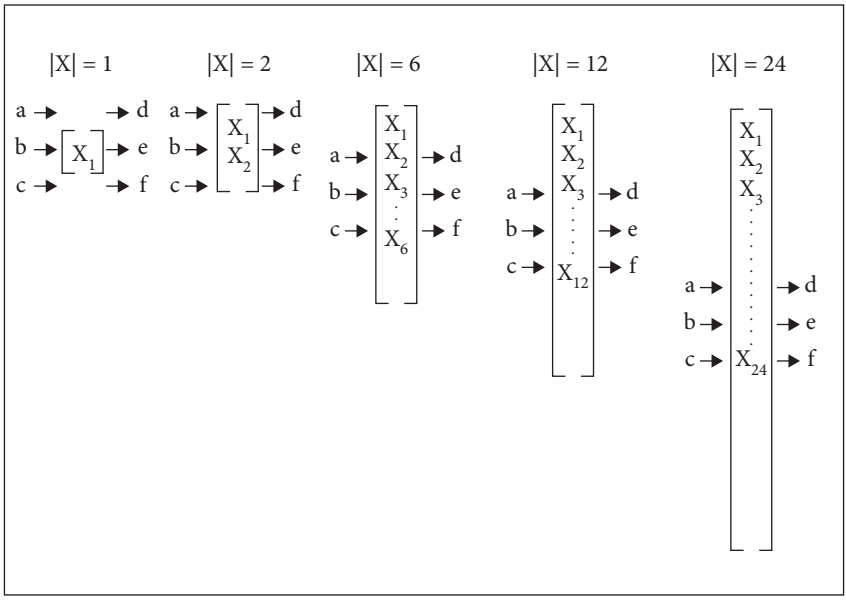

Figure 1. The miniature grammars used by Gómez (2002; columns 2–5) and Onnis et al. (2003; submitted; columns 1–5). Sentences with three non-adjacent dependencies are constructed with an increasing number of syntagmatically intervening X items. Gómez used set-sizes 2, 6, 12, and 24. Onnis et al. added a new set-size 1 condition

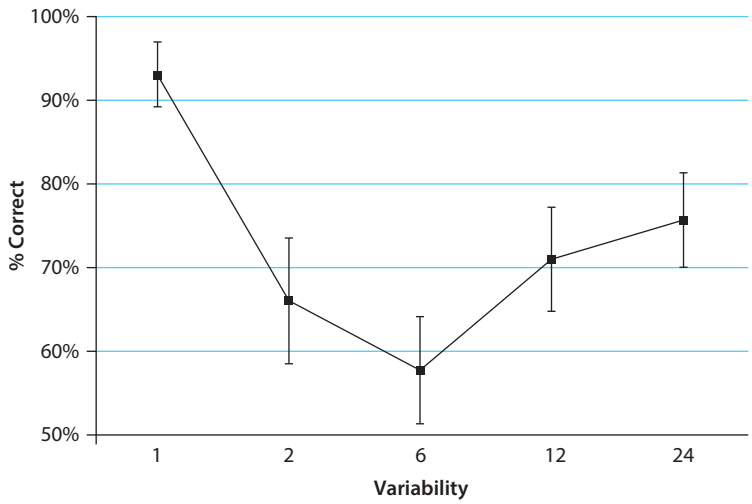

Figure 2. Data from Onnis et al. (2003, submitted) incorporating the original Gomez experiment. Learning of non-adjacent dependencies results in a U shape curve as a function of the variability of intervening items, in five conditions of increasing variability

Candidate measures of associative learning

There exist several putative associative mechanisms of artificial grammar and sequence learning (e.g. Dulany et al. 1984; Perruchet & Pacteau, 1990; Servan-Schreiber & Anderson, 1990), or on learning of whole items (Vokey & Brooks, 1992). Essentially these models propose that subjects acquire knowledge of fragments, chunks or whole items from the training strings, and that they base their subsequent judgments of correctness (grammaticality) of a new set of sequences on an assessment of the extent to which the test strings are similar to the training strings (e.g. how many chunks a test item shares with the training strings). To find out how well these associative models would fare in accounting for Gómez and for Onnis et al.'s data, we considered a variety of existing measures of chunk strength and of the similarity between training and test exemplars. Based on existing literature, we considered the following measures: Global Associative Chunk Strength (GCS), Anchor Strength (AS), Novelty Strength (NS), Novel Fragment Position (NFP), and Global Similarity (GS), in relation to the data in Experiment 1 and 2 of Onnis et al. These measures are described in detail in Appendix A. Table 1 summarizes descriptive fragment statistics are summarized, while the values of each associative measure are reported in Table 2.

Table 1. Descriptive fragment statistics for the bigrams and trigrams contained in the artificial grammar used in Gómez (2002), Experiment 1, and in Onnis et al. (submitted). Note that Experiment 1 of Onnis et al. is a replication of Gómez' (2003) Experiment 1

Variability condition	1	1-cntrl	2	6	12	24
Total number of training strings	432	432	432	432	432	432
A_i–B_i pair types	3	6	3	3	3	3
A_i–B_i pair tokens	144	72	144	144	144	144
X_j types	1	1	2	6	12	24
X_j tokens	432	432	216	72	36	18
$A_iX_jB_i$ types	3	6	6	18	36	72
$A_iX_jB_i$ tokens	144	72	72	24	12	6
type/token ratio (AXB)	0.02	0.08	0.08	0.75	3.00	12.00
A_iX_j tokens	144	72	72	24	12	6
X_jB_i tokens	144	72	72	24	12	6
$P(X_j\|A_i)$	1.00	1.00	0.50	0.17	0.08	0.04
$P(B_i\|X_j)$	0.33	0.16	0.33	0.33	0.33	0.33

Table 2. Predictors of chunk strength and similarity used in the AGL literature (Global Chunk Strength, Anchor Chunk Strength, Novelty, Novel Fragment Position, Global Similarity). Scores refer to bigrams and trigrams contained in the artificial grammar used in Gómez (2002), Experiment 1, and Onnis et al. (submitted)

Variability condition	1	1-cntrl	2	6	12	24
GCS/ACS for Grammatical strings	144	72	72	24	12	6
GCS/ACS for Ungrammatical strings	96	48	48	16	8	4
Novelty for Grammatical strings	0	0	0	0	0	0
Novelty for Ungrammatical strings	1	1	1	1	1	1
NFP for Grammatical strings	0	0	0	0	0	0
NFP for Ungrammatical strings	0	0	0	0	0	0
GS for Grammatical strings	0	0	0	0	0	0
GS for Ungrammatical strings	1	1	1	1	1	1

The condition of null variability (set-size 1) is the only condition that can a priori be accommodated by measures of associative strength. For this reason, the set-size 1-control was run in Experiment 2. Table 2 shows that associative measures are the same for the set-size 1-control and set-size 2. However, since performance was significantly better in the set-size 1-control, the above associative measures cannot predict this difference.

Overall, since Novelty, Novel Fragment Position, and Global Similarity values are constant across conditions, they predict that learners would fare equally in all conditions and, to the extent that ungrammatical items were never seen as whole strings during training, that grammatical strings would be easier to recognize across conditions. Taken together, the predictors based on strength and similarity would predict equal performance across conditions or better performance when the set-size of embeddings is small because the co-occurrence strength of adjacent elements is stronger. Hence, none of these implicit learning measures predict the observed U shape results. In the next section, we investigate whether connectionist networks can do better, and whether any particular network architecture is best.

Simulation 1: Simple recurrent networks

We have seen that no existing chunk-based model derived from the implicit learning literature appears to capture the U-shaped pattern of performance exhibited by human subjects when trained under conditions of differing variability. Would connectionist models fare better in accounting for these data? One plausible candidate is the Simple

Recurrent Network model (Elman, 1990) because it has been applied successfully to model human sequential behavior in a wide variety of tasks including everyday routine performance (Botvinick & Plaut, 2004), dynamic decision making (Gibson, Fichman, & Plaut, 1997), cognitive development (Munakata, McClelland, & Siegler, 1997), implicit learning (Kinder & Shanks, 2001; Servan-Schreiber, Cleeremans, & McClelland, 1991), and the high-variability condition of the Gómez (2002) nonadjacency learning paradigm (Misyak et al. 2010b). SRNs have also been applied to language processing such as spoken word comprehension and production (Christiansen, Allen, & Seidenberg, 1998; Cottrell & Plunkett, 1995; Dell, Juliano, & Govindjee, 1993; Gaskell, Hare, & Marslen-Wilson, 1995; Plaut & Kello, 1999), sentence processing (Allen & Seidenberg, 1999; Christiansen & Chater, 1999; Christiansen & MacDonald, 2009; Rohde & Plaut, 1999), sentence generation (Takac, Benuskova, & Knott, 2012), lexical semantics (Moss, Hare, Day, & Tyler, 1994), reading (Pacton, Perruchet, Fayol, & Cleeremans, 2001), hierarchical structure (Hinoshita, Arie, Tani, Okuno, & Ogata, 2011), nested and cross-serial dependencies (Kirov & Frank, 2012), grammar and recursion (Miikkulainen & Mayberry III, 1999; Tabor, 2011), phrase and syntactic parsing (Socher, Manning, & Ng, 2010), and syntactic systematicity (Brakel Frank, 2009; Farkaš & Croker, 2008; Frank, in press). In addition, recurrent neural networks are effectively solve a variety of linguistic engineering problems like automatic voice recognition (Si, Xu, Zhang, Pan, & Yan, 2012), word recognition (Frinken, Fischer, Manmatha, & Bunke, 2012), text generation (Sutskever, Martens, & Hinton, 2011), and recognition of sign language (Maraqa, Al-Zboun, Dhyabat, & Zitar, 2012). Thus these networks are potentially apt at modeling the difficult task of learning of non-adjacencies in the *AXB* artificial language discussed above. In particular, SRNs (Figure 3a) are appealing because they come equipped with a pool of units that are used to represent the temporal context by holding a copy of the hidden units' activation level at the previous time slice. In addition, they can maintain simultaneous *overlapping*, *graded* representations for different types of knowledge. The gradedness of representations may in fact be the key to learning non-adjacencies. The specific challenge for SRNs in this paper is to show that they can represent graded knowledge of bigrams, trigrams and non-adjacencies and that the strength of each such representation is modulated by the variability of embeddings in a similar way to humans.

To find out whether associative learning mechanisms can explain the variability effect, we trained SRNs to predict each element of the sequences that were structurally identical to Gómez's material. The choice of the SRN architecture, as opposed to a simple feed-forward network, is motivated by the need to simulate the training and test procedure used by Gómez and Onnis et al. who exposed their participants to auditory stimuli, one word at a time. The SRN captures this temporal aspect.

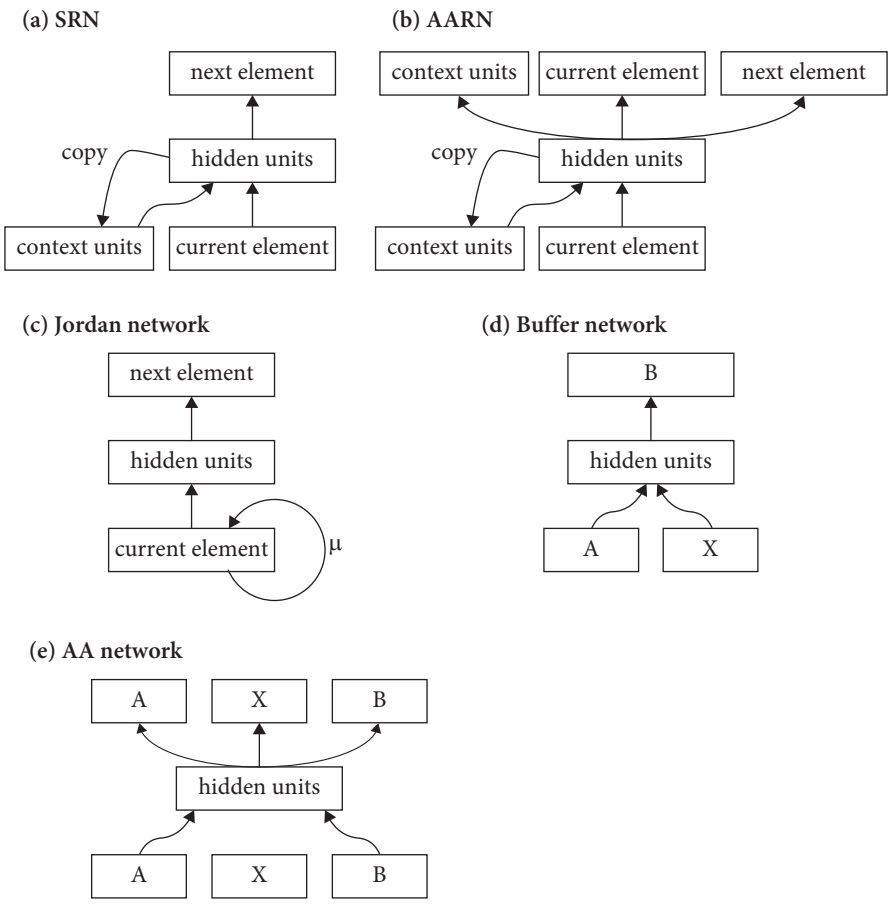

Figure 3. Network architectures tested in Simulations 1 and 2

Method

Network architecture and parameters

SRNs with 5, 10, and 15 hidden units and localist representations[1] on the input and output units were trained using backpropagation on the strings designed by Gómez. For each of the three hidden unit variations of the SRN, we systematically manipulated 5 values of learning rate (0.1, 0.3, 0.5, 0.7, 0.9) and five values of momentum (0.1, 0.3, 0.5, 0.7, 0.9). Each network was initialized with different random weights to simulate a

1. Each word was an input vector with all units set to zero and a specific unit set to 1.

different participant. Learning rate, momentum, and weight initialization were treated as corresponding to individual differences in learning in the human experiments, where indeed some considerable variation in performance was noted within variability conditions. Strings were presented one element at a time to the networks by activating the corresponding input unit. Thirty-one input/output units represented the three initial (A_i) elements, the three final (B_i) elements, one of the 24 possible embedded (X_j) elements, and an End-of-String marker. Gómez and Onnis et al. used longer pauses between the last word of a string and the first word of the following strings, to make each three-word string perceptually independent. Similarly, the End-of-String marker informed the networks that a new separate string will follow, and context units were reset to 0.0 after each complete string presentation. On each trial, the network had to predict the next element of the string, and the error between its prediction and the actual successor to the current element was used to modify the weights.

Materials

Both training and test stimuli consisted of the set of strings used in Onnis et al.'s Experiment 1, which incorporated Gómez's Experiment 1 and added the zero-variability condition.[2] During training, all networks were exposed to the same total number of strings (1080 strings, versus 432 in Gómez's experiment),[3] so that each would experience exactly the same number of non-adjacencies. This required varying the number of times the training set for a particular variability condition was presented to the network. Thus, while in the set-size 24 condition the networks were exposed to 15 repetitions of the 72 possible string types, in the set-size 2 condition they were exposed to 180 repetitions of the 6 possible string types.

Procedure

Twenty networks × 5 conditions of variability × 3 hidden-unit × 5 learning-rate × 5 momentum parameter manipulations were trained, resulting in 7500 individual networks being trained, each with initial random weights in the –0.5, +0.5 range. After training, the networks were exposed to 12 strings, 6 of which were part of the trained language in all set-size conditions, and 6 of which were part of a novel language in which the pairings between initial and final elements had been reversed so that each

2. Gómez used two languages where the end-items were cross-balanced to control for potential confounds. Because our word vectors are orthogonal to each other, we created and tested only 1 language.

3. This value was determined empirically so as to produce good learning in the MacIntosh version of the PDP simulator with the parameters we selected. Typically neural networks require a longer training – tens of thousand epochs – to start reduce their error. Thus a training of 1080 epochs, although longer than the human experiment, is a reasonably close approximation to 432.

head was now associated with a different final element. Test stimuli consisted of 3 grammatical strings and 3 ungrammatical strings repeated twice, as in Onnis et al.[4] The large parameter manipulations were motivated by the need to test the robustness of the findings.

Network analysis

Networks were tested on a prediction task. Performance was measured as the relative strength of the networks' prediction of the tail element *B* of each *AXB* sentence when presented with its middle element *X*. The activation of the corresponding output unit was recorded and transformed into Luce ratios (Luce, 1963) by dividing it by the sum of the activations of all output units:

$$Luce \quad \frac{output_{target}}{\sum output}$$

Luce ratios were calculated for both grammatical and ungrammatical test strings. Good performance occurred when Luce ratios for grammatical strings (e.g. A_iXB_i) were high, i.e. showing an ability to activate the correct target output unit, while Luce ratios for ungrammatical strings (e.g. A_iXB_j) were close to zero. This is captured by a high value of Luce activation differences between grammatical and ungrammatical activation values. If the networks did not learn the correct non-adjacent pairs, either all three target output units for the *B* item would be equally activated when an *X* was presented – resulting in a value close to zero for Luce ratio differences, or typically only one wrong non-adjacent dependency would be learned, as a result of the networks finding a local minimum – in which case Luce ratio differences would still be close to zero.

Results and discussion

Luce ratio values were averaged over the 20 replications in each condition, and across learning rate and momentum conditions for each of the 3 hidden unit variations of SRNs. To directly compare the networks results with human data we computed z-scores of Luce ratio differences between grammatical and ungrammatical responses for each network and z-scores of differences between correct incorrect raw score differences for each participant in Onnis et al. As can be seen in Figure 4 the best candidate networks that reproduced the U shape closer to the human data had 10 hidden

4. Given that in set-size 1 humans and networks are trained on a single embedding they could only be tested on strings containing one embedding. Hence networks were tested on 6 strings repeated twice.

units (Figure 4). These results provide two findings: firstly, there is at least one class of associative learning machines implemented in SRNs that are able to learn nonadjacent dependencies. Secondly, there is at least one class of associative learning machines implemented in SRNs that learn nonadjacencies in a similar way to humans, i.e. with performance being a U-shaped function of the variability of intervening items.

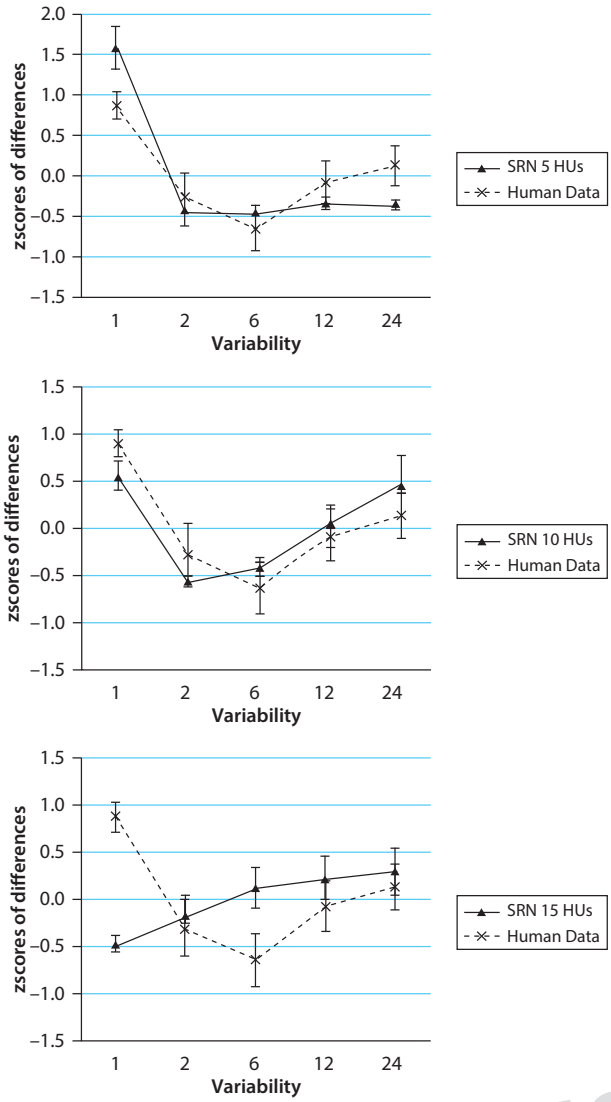

Figure 4. Comparison between SRNs with 5, 10, and 15 hidden units (hu) and human data (HD). The SRNs with 10 hidden units provide the best match with participants' performance

Can other connectionist architectures capture the data?

The motivation for using SRNs in Simulation 1 is based on the wide type of sequential behaviors they can capture as evidenced in the literature (see references above). However, other well-known architectures such as the Jordan Network and the Auto-Associative Recurrent Network share many features with the SRN, in particular they also incorporate mechanisms to represent time via recurrent connections. In the simulations below we trained and tested four different types of connectionist networks on the variability task: Auto-Associative Networks (AAN), Jordan Networks, Buffer Networks and Auto Associators (AA). Notably, all network architectures were trained with the same training regime and parameter manipulations as the SRNs, and their performance was measured in terms of normalized Luce ratio differences, thus allowing direct comparison with both the SRNs in Simulation 1, and the human data. Below we present four Methods sections separately, each corresponding to the four net architectures. A single Results section will then directly compare the four architectures' performance.

Simulation 2a: Auto-associative recurrent networks

The Auto-Associative Recurrent Network (henceforth, AARN) has been proposed by Maskara and Noetzel (1992; see also Dienes, 1992). The AARN is illustrated in Figure 3b. As its name suggests, this network is essentially an SRN that is also required to act as an encoder on both the current element and the context information. On each time step, the network is thus required to produce the current element and the context information in addition to predicting the next element of the sequence. This requirement forces the network to maintain information about the previously presented sequence elements that would tend to be "forgotten" by a standard SRN performing only the prediction task. Maskara and Noetzel showed that the AARN is capable of mastering languages that the SRN cannot master.

Method

Twenty AARNs with different random weights × 3 hidden unit × 5 variability condition × 5 learning rates × 5 momentum manipulations for a total of 7500 separate simulations were trained and tested with exactly the same training and test regime and strings as the SRN. Performance of the AARN was assessed in exactly the same way as it was done for the SRN. In the test phase, when presented with the middle element of each sequence, we compared the activation of the target unit in the output units corresponding to the tail B_i element for the grammatical and ungrammatical sequences.

Simulation 2b: Jordan networks

Jordan Networks (Jordan, 1986, Figure 3c) assume that the recurrent connections that make them sensitive to temporal relationships possible in the SRN occur not between hidden and context units, but between output units and state units. Thus, on each time step, the network's previous output is *blended* with the new input in a proportion defined by a single parameter, μ. The parameter is used to perform time-averaging on successive inputs. This simple mechanism makes it possible for the network to become sensitive to temporal relationships because distinct sequences of successive inputs will tend to result in distinct, time-averaged input patterns (within the constraints set by the simple, linear time-averaging). However, it should be clear that the temporal resolution of such networks is limited, to the extent that the network, unlike the SRN, never actually has to learn how to *represent* different sequences of events, but instead simply relies on the temporally "pre-formatted" information made possible by the time-averaging. In Jordan's original characterization of this architecture, the network's input units also contained a pool of so-called "plan" units, which could be used to represent entire subsequences of to-be-produced outputs in a compact form. Such "plan" units have no purpose in the simulations we describe, and were therefore not incorporated in the architecture of the network.

Method

Twenty Jordan nets × 3 hidden unit × 5 variability conditions × 5 learning rates × 5 momentum manipulations resulted in 7500 different simulations being trained and tested with exactly the same training and test regime and strings as the SRN. The μ parameter was set to 0.5. Performance was again assessed in the same way as for the SRN. In the test phase, upon presentation of a middle element *X*, the level of activation of the target unit of the pool of output units corresponding to the tail B_i element was compared for the grammatical and ungrammatical sequences.

Simulation 2c: Buffer networks

Buffer Networks (Figure 3d) are three-layer feed-forward networks in which pools of input units are used to represent inputs that occur at different time steps. On each time step during the presentation of a sequence of elements, the contents of each pool are copied (and possibly decayed) to the one that corresponds to the previous step in time, and a new element is presented on the pool that corresponds to time *t*, the current time step. The contents of the pool corresponding to the most remote time

step are discarded. Because of its architecture, the buffer network's capacity to learn about temporal relationships is necessarily limited by the size of its temporal window. In our implementation of the buffer architecture, the task of the network is to predict the third element of a sequence based on the first and second elements. The size of the temporal window is therefore naturally limited to two elements of temporal context. Thirty units were used to represent both initial and middle elements (six initial/final elements and 24 possible middle elements). The task of the network was to predict the identity of the final element of each sequence. Six output units, corresponding to the six A_i and B_i items, were used to represent the final element.

Method

Twenty Buffer nets × 3 hidden unit × 5 variability conditions × 5 learning rates × 5 momentum × 2 decay parameter manipulations resulted in 15000 different simulations being trained and tested with exactly the same training and test regime and strings as the SRN. Decay parameters were 0.0 and 0.5. Performance was again assessed in the same way as it was done for the SRN and the other nets.

Simulation 2d: Auto-associator networks

The task of the Auto-Associator network (Figure 3e) simply consists in reproducing at the output level the pattern presented at the input level. In our implementation, the entire three elements strings were presented at the same time to the network by activating three out of thirty input units corresponding to the initial, middle and final elements of each sequence. Performance was assessed in the test phase by comparing, between grammatical and ungrammatical strings, the level of activation of the target output unit corresponding to the final element.

Method

Twenty AA nets × 3 hidden unit manipulations × 5 variability conditions × 5 learning rates × 5 momentums resulted in 7500 different simulations. Training and test procedures were exactly the same as for the previous network simulations.

Results

All results (Figure 5) are plotted as z-score transformed values of Luce ratio differences between network predictions for grammatical and ungrammatical test strings. Note that these are *average* z-score values across *all* different parameter manipulations for

each network architecture. AARNs perform very poorly on set-size 1, while learning fairly well but equally across all other conditions. This pattern of results is difficult to interpret in the light of the human data (indicated as HD in Figure 5).

Jordan nets show a pattern similar to the AARN network, with very poor performance at set-size 1 and best performance at set-size 2, which almost reverses the pattern of behavioral data.

Buffer nets draw a steady curve above 0 for zero and small set-size conditions, indicating some moderate but equal learning across such conditions. Performance descends abruptly for set-size 24. This pattern of results also fails to replicate the

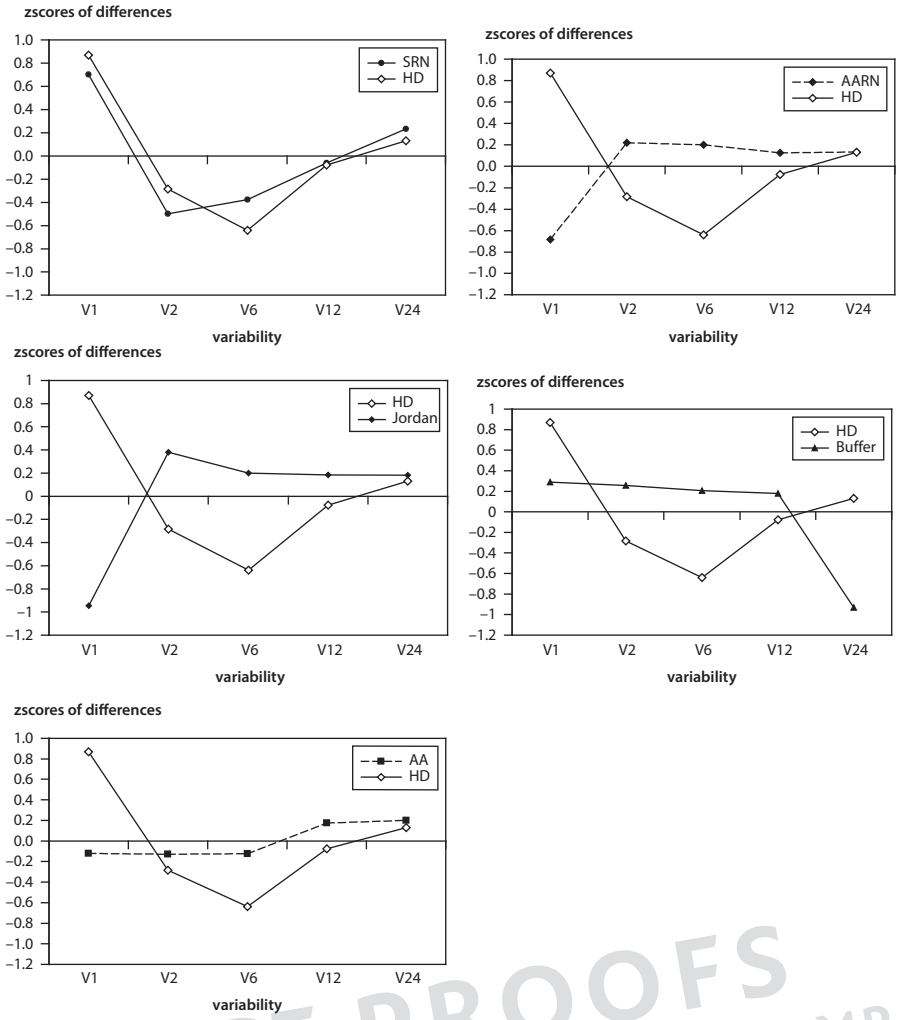

Figure 5. Comparison between the 5 connectionist architectures and human data (HD)

human U shape. Rather it seems to suggest that performance is based on the strength of local bigram and trigram information.

Finally, AA networks draw a relatively flat curve across conditions, which again does not replicated the U shape. Rather it shows that, on average, the AA is not influenced by the variability of the middle element.

Figure 5 visualizes an *average* performance of the different architectures when z-score values are averaged across all different parameter and hidden unit manipulations. However, averaging might conflate parameter configurations that work really well with others that work very poorly. Similarly, the SRN parameter configuration that we have chosen might stand alone among the class of SRNs able to capture the U shape. The robustness of a specific architecture can be visualized by assessing how densely several configuration simulations of the same architecture inhabit the region of space corresponding to good performance. A 2-dimensional graph (Figure 6) was plotted where, for each network parameter configuration, the *x* axis plotted z-score differences between performance at set-size 24 and the mean performance of set-sizes 2, 6, and 12, whereas the *y* axis plotted z-score differences between performance at set-size 1 and the mean performance of set-sizes 2, 6, and 12. The graph splits into 4 different quadrants, divided at 0 both on the *x* and the *y* axis. This graph captures the U-shaped nature of the behavioral data: if performance is good on both set-sizes 1 and 24, while being poor at the same time on set-sizes 2, 6, and 12, both *x* and *y* z-scores will be higher than 0, falling in the upper right quadrant of the graph. This is indeed where the human data are located. We produced five such graphs for the 5 connectionist architectures. From the graphs one can see that the SRN is the architecture closest to the human data, regardless of parameter variations, although performance is better with largest variability than with no variability. Conversely, most AARN simulations cluster in the lower left quadrant, indicating that performance at both endpoints of variability tends to be lower than with small variability. The Jordan nets follow a similar, sparser trajectory, whereas Buffer nets are able to learn in set-size 1 conditions but fail at set-size 24. Lastly, AA nets display virtually no variation due to parameter manipulation and cluster tightly at the exact intersection of the four quadrants, indicating that they tend to learn equally well in all 5 set-size conditions.

To summarize, having compared five different connectionist architectures against the human data, we can conclude that the SRN is the connectionist model that best captures such data. This result is especially interesting considering that Jordan nets and AARN nets belong to the same class of recurrent nets and that the AARN was proposed as a better alternative architecture to the SRN. But what specific aspect of the SRN allows it to fit the data best? Next we probe the hidden units, which carry the internal representations of the network, for possible clues to this question.

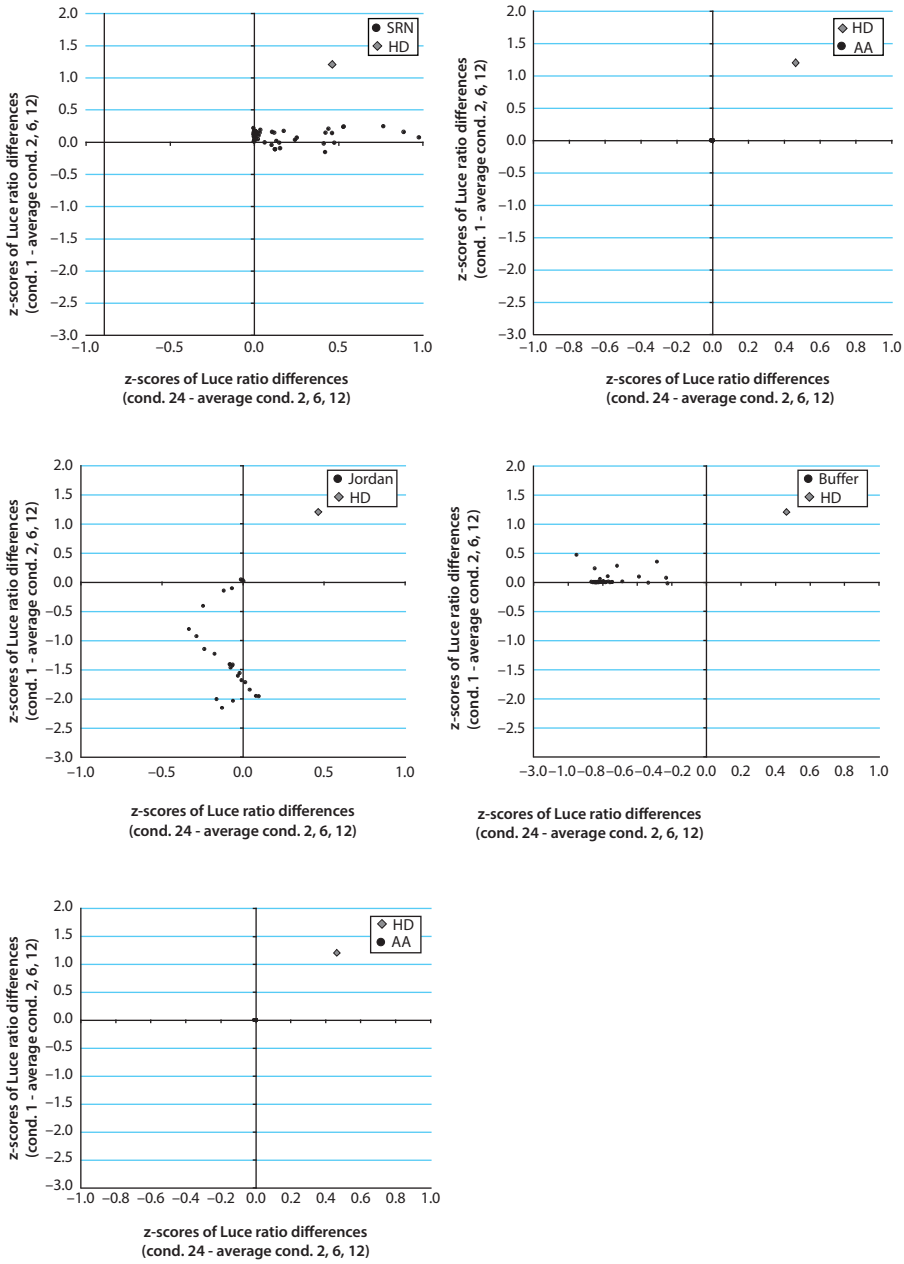

Figure 6. The regions of the space inhabited by the 5 connectionist architectures. Only SRNs group in the upper-right quadrant, where human data from Onnis et al. (2003; submitted) are located

Mechanisms of implicit learning in the SRN

Simulation 1 gathered evidence that SRNs trained to predict each element of sequences identical to those used in Gómez (2002) and Onnis et al. (2003, submitted) can master non-adjacencies in a manner that depends on the variability of the intervening material, thus replicating the empirically observed U-shaped relationship between variability and classification performance. The specific interest paid to the U shape results in this study lies in the fact that no mechanism proposed in the implicit learning literature can readily simulate the human data. To the extent that SRNs are also associative machines, the successful results are also surprising. In this section we attempt to understand how SRNs succeed in learning non-adjacencies.

The key to understanding the SRN behavior is its ability to represent in its hidden units *graded* and *overlapping* representations for both the current stimulus-response mapping and any previous context information. Hidden units adjust at each step of processing and can be thought of as a compressed and context-dependent "re-representation" of the current step in the task. Given for instance a network with 10 hidden units, the internal representation of this network can be seen as a point in a 10-dimensional space. As the training progresses, the network's representation changes, and a trajectory is traced through the 10-dimension space. Multi Dimensional Scaling (MDS) is a technique that reduces an n-dimensional space into a 2-dimensional space of relevant dimensions, and thus allows the visualization of this learning trajectory in a network as a function of training (Figures 7, 8, and 9). In order to predict three different B_i endings correctly the network has to develop trajectories that are *separate enough* at the time that an X_j is presented (see also Botvinick & Plaut, 2004).

For the sake of the argument, let us consider a simpler scenario in which an artificial language is composed of only two items, i.e. it is an X_jB_i language. When the input is an *X,* the hidden units must be shaped so as to predict one of three *B* elements. This task still requires some considerable learning because the net has to activate an output node out of all the possible items in the language, including the *X*s. What specific *B* will they predict? The hidden units are modified by both (a) a trace for each of the *X*s from the input units at time *t*, and (b) the EOS (End of Sentence) marker from the context units (this was information at time *t*-1). In this case, given that this past information is exactly *identical* for whatever prediction of *B*, the hidden unit representations will be similar regardless of any specific B_i continuation. In this case, therefore, there is absolutely no information in the past items that can help the hidden units to develop *separate* trajectories for B_1, B_2, and B_3, and the best error reduction is obtained by activating the nodes corresponding to the three *B*s with an activation of 0.33 (corresponding to an even probability of predicting one of three elements).

Let us now imagine the scenario of our simulations in which the language is an A_iXB_i language. Here the past information that shapes the hidden units is (a) a trace from one of several *X*s from the input units at time *t*; (b) a trace from one of three *A*s at time *t*-1 from the context units, which is specific for each *B* prediction; and (c) a trace from the previous EOS (End of Sentence) marker which has been incorporated in the previous time steps at *t*-2, which is the same for all *B* predictions. The past context for predicting a specific B_i is now partially different, because we have a specific correspondence between an A_i and a B_i in the language. In this scenario the hidden units *may* now develop different trajectories, and thus be able to predict successfully different *B* continuations. What is the best condition for such dissimilarity? With low variability of *X*s the traces from each shared *X* overshadow the traces from the *A* elements so that the networks form very similar representations for predicting *B* elements. Figure 7 presents the two principal components of a Multiple Dimensional Scaling (MDS) analysis over the SRN hidden units in the setsize 2 condition, at the time of predicting the *B* element over 15 different points in training.[5] Hidden unit trajectories move across training, but they do not separate at the end of training. Contrast this result with Figure 8, the same MDS analysis over the hidden units of a network in Set-size 24. Hidden units move together in space at the beginning of training up to a point when they separate in 3 different sub-regions of the space, corresponding to 3 separate representations for A_1, A_2, and A_3. It is evident that the 24 embeddings now each contribute a *weaker* trace and this allows the trace from each individual A_i element to be maintained more strongly in the context units, shaping the activation pattern of hidden units.

Regarding the large difference in performance between set-size 1 and 2, how do SRNs learn to predict the correct *B* non-adjacency in the former but not in the latter case? The MDS graph of hidden unit trajectories (Figure 9) once again reveals that different trajectories are traversed ending in three distinct regions of the space, a situation similar to set-size 24. It seems that the networks develop a compressed representation for a general *X* either with no variability or with a large enough number of *X*s, thus leaving computational space for the three distinct *A* traces to be encoded in the hidden units.

Although this explanation is reasonable for set-size 24, one possibility is however that the networks merely memorize the three different strings in set-size 1, suggesting that not one but two different mechanisms are responsible for the U shape – one based on variability in set-size 24 and one based on rote learning in set-size 1. In Onnis et al.

5. Ungrammatical sequences are removed from the graph, because each produces exactly the same vector over the network's hidden units. Hence the graph displays 6 trajectories: one each for AX_1, AX_2, BX_1, BX_2, CX_1, and CX_2.

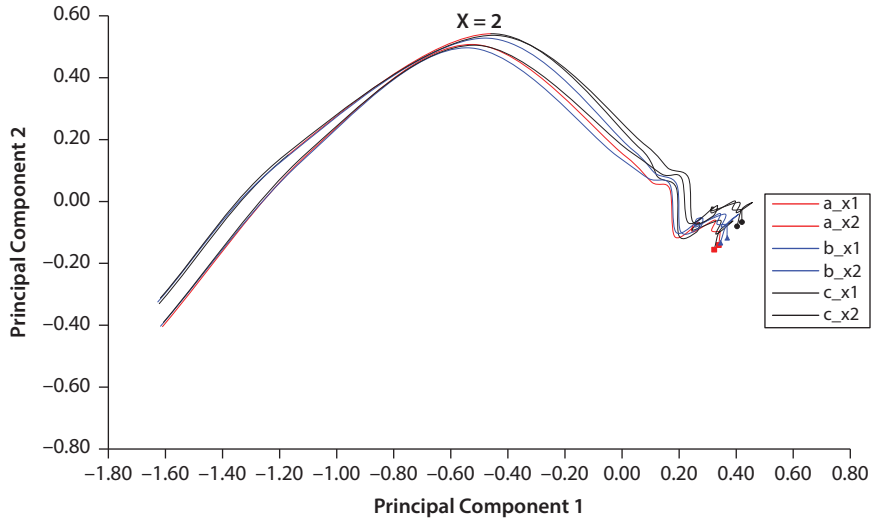

Figure 7. MDS analysis of hidden unit trajectories. A network trained on 2 Xs fails to achieve the needed separation: all 6 trajectories remain close to each other all the way through the end of training. Hence the network can never form correct predictions of the successor to the X

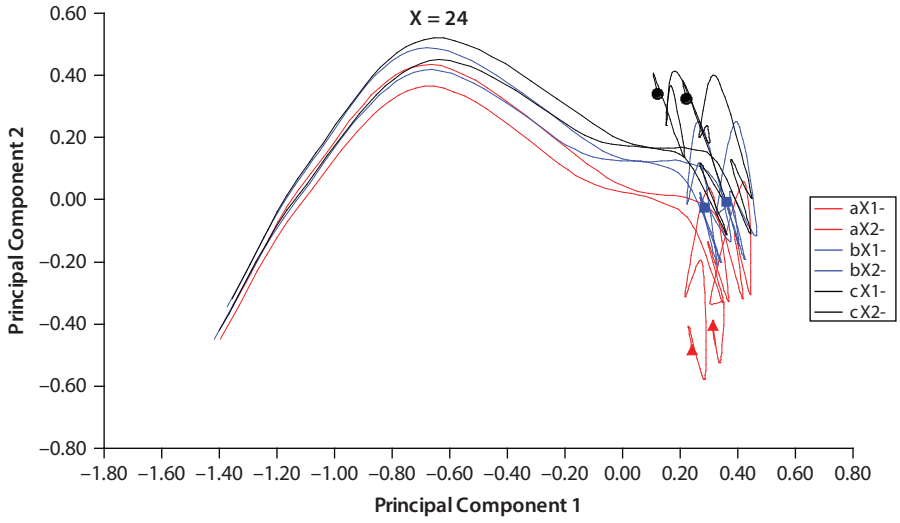

Figure 8. MDS analysis of hidden unit trajectories in the set-size 24 X condition: all 6 trajectories start out, on the left side, from the same small region, and progressively diverge to result in three pairs of two representations

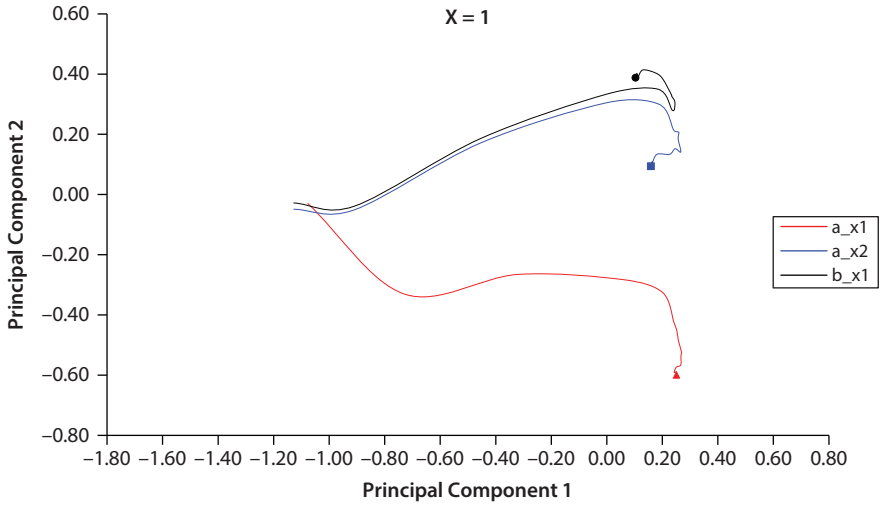

Figure 9. MDS analysis for a network trained on the set-size 1 X condition. Like in the set-size 24 X case, the network is successful in separating out the corresponding internal representations: The terminal points of each trajectory end up in different regions of space

this possibility was resolved by showing that learners can endorse correct nonadjacent dependencies in set-size 1 even when presented with a novel *X* at test (their Experiment 3). They also showed (Experiment 2) that performance was good even when 6 different *A_B* pairs had to be learned with one *X*. Since in this latter control condition the number of string types to be learned was exactly the same as in set-size 2 (and indeed resulted in a more complex language with 13 words as opposed to 7 words in set-size 2), the difference in performance could not be accounted for by a memory advantage in set-size 1. Since the MDS analyses cannot disambiguate whether the networks learn by rote in set-size 1 – a result that would differ from human learning – we ran further simulations equivalent to Onnis et al.'s Experiment 2 and 3. SRNs were trained on exactly the same training regime as Simulation 1, while $A_i\_B_i$ and $^*A_i\_B_j$ frames were presented at test with a completely new *X* that had never appeared during training. Intriguingly, the networks still recognized the correct non-adjacencies better with null or high variability than in the set-size 2 condition. Figure 10 shows that when presented with novel *X*s at test SRNs performance is considerably better in set-size 1 and 24 than in set-size 2. Figure 11 shows that this advantage persists when the networks have to learn 6 nonadjacent dependencies, i.e. when the number of trigrams to be learned is equated in set-size 1 and set-size 2. Crucially, in both set-size 1 and 24, the networks develop a single representation for the *X*, which leaves compression space for the trace of distant *A* elements to be encoded in the hidden units.

We believe that these results, coupled with the separation of hidden unit trajectories, form compelling evidence that the learning of non-adjacencies happens

independently of specific *X* embeddings, thus corroborating the idea that what is learned is not a trigram sequence of adjacent elements, but a true discontinuity relation. In fact, the reason why the discontinuities are not learned equally well in low variability conditions is exactly that the networks find an optimal solution in learning adjacent bigram information in those conditions. Our simulations reveal that a similar variability-driven mechanism is responsible for better learning of non-adjacencies in either zero or high variability, closely matching the human data.

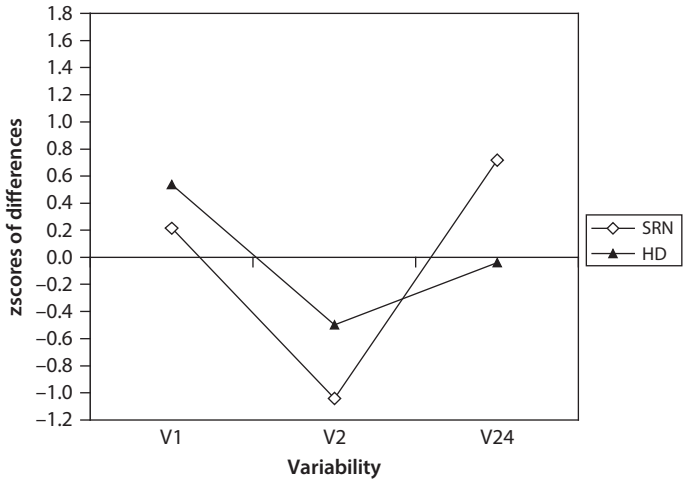

Figure 10. SRN and Human Data (HD) performance in endorsing nonadjacencies in sentences containing novel Xs in three differing conditions of variability (V)

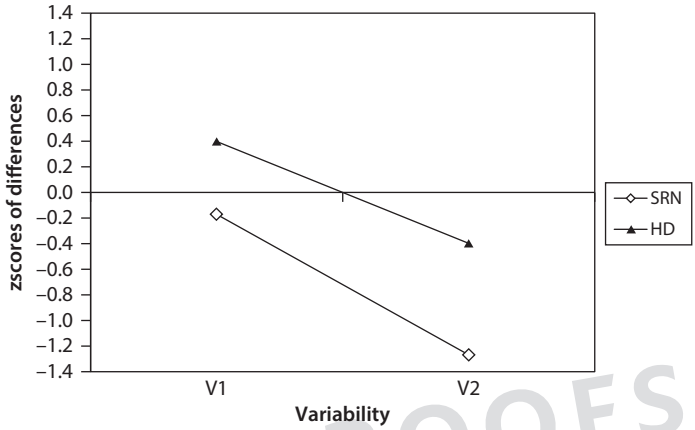

Figure 11. Both SRNs and humans learn better six nonadjacent frames with one X than three nonadjacent frames with two Xs, suggesting that there is something special about having only one intervening X

Type variability or token frequency?

A last confound that has to be disentangled in the current simulations is the possible role of token frequency of *X* elements. Because the total number of learning trials is kept constant across conditions, in larger set-size conditions each *X* element is presented to the network fewer times. It may thus be the case that the trace from the *A* elements can be better encoded in the hidden units in set-size 24 because the token frequency of each *X* element decreases. Under this scenario, improved non-adjacent learning in higher-variability conditions would not necessarily be due to the higher variability of *X types*, but rather to the lower frequency of *X tokens*, thus perhaps trivializing our results.

Therefore, we ran a further set of simulations similar to Gómez (2002), in which the number of total token presentations of each *X* was kept constant across conditions. Gómez found that learning still improved in set-size 24, thus ruling out the impact of *X* token frequency. Figure 12 shows that when the number of *X* tokens is identical across variability conditions to the one used in set-size 24 condition of Simulation 1 (i.e. 15 repetitions) then the SNRs learn in the high variability conditions, suggesting that type variability, not token frequency is indeed the key factor improving performance in set-size 24. Figure 12 also shows that with training brought up to more asymptotical levels (token frequency of *Xs* of 150, 360, and 720 repetitions held constant across set-size conditions) the U shape is restored. These training trajectories are in line with connectionist networks' typical behavior and do not depart from human behavior. Typically a connectionist network needs a certain amount of training in order to get "off the ground". It starts with low random weights, and needs to configure itself to solve the task at hand. This takes several training items, many more than humans typically need. Arguably when humans enter the psychologist's lab to participate in a study they do not start with "random connections", rather they bring with them considerable knowledge, accumulated over years of experience with sequences of events in the world. Therefore, we expected networks to require a longer training to configure themselves for a particular task. In separate studies, connectionist networks were pre-trained on basic low-level regularities of the training stimuli prior to the actual learning task (Botvinick & Plaut, 2006; Christiansen, Conway, and Curtin, 2000; Destrebecqz & Cleeremans, 2003; Harm & Seidenberg, 1999). As more data is collected on the learning of non-adjacencies, it will be necessary to provide more detailed models. However, our choice of localist representations and no pre-training was motivated by the desire to capture something general about the U shape, as Onnis et al. (submitted; Experiment 4) also obtained a similar learning with visually presented pseudo-shapes. Therefore, Figure 12 suggests that when the SRNs receive sufficient training to learn the material in every condition (at least 150 repetitions of each *X* element) the U-shaped curve is fully restored. These control simulations

suggest that the emerging U-shaped curve in learning non-adjacencies is truly mediated by the type frequency of intervening embedded elements.

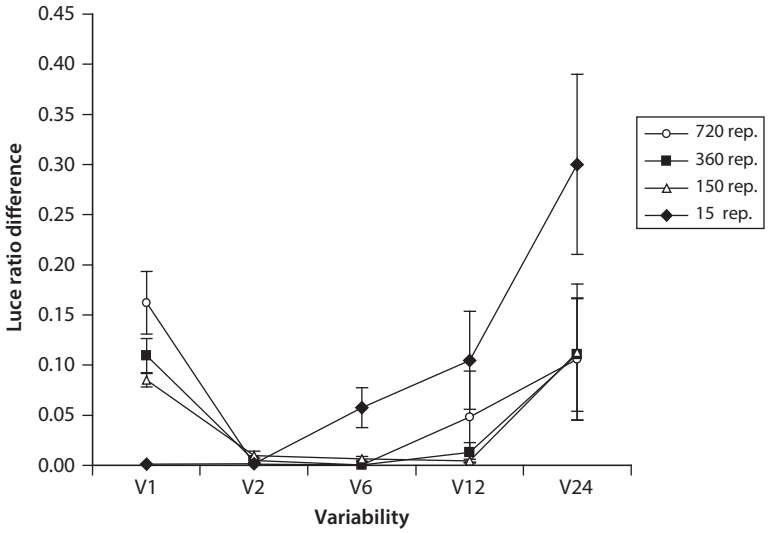

Figure 12. SRNs simulations controlling for number of tokens of the embeddings across variability conditions. With a sufficient number of tokens (150) the networks display a U shape learning curve that is dependent on the variability of embeddings

Conclusions

Sensitivity to transitional probabilities of various orders including non-adjacent probabilities in implicit sequential learning has been observed experimentally in adults and children, suggesting that learners exploit these statistical properties of the input to detect structure. Indeed, studies of individual differences in the ability to detect nonadjacencies in implicit sequential learning tasks have been found to correlate with adults' language skills (Misyak & Christiansen, 2012; Misyak et al. 2010a, b). Detecting non-adjacent structure poses a genuine computational and representational problem for simple associative models based purely on knowledge of adjacent items. Following Gómez (2002), a more elaborate proposal is that human learners may exploit different sources of information, here adjacencies and non-adjacencies, to learn structured sequences. Her original results suggested that non-adjacencies are learned better when adjacent information becomes less informative.

The current work began where the experimental data of Gómez (2002) and Onnis et al. (2003; submitted) concluded. It is a first attempt to provide a mechanistic account of implicit associative learning for a set of human results that the current literature

cannot explain. We have compared 5 different connectionist architectures with several different parameter configurations resulting in 22,500 individual simulations, allowing a comprehensive search over the space of possible network performances. Such extensive modelling allowed us to select with a good degree of confidence Simple Recurrent Networks as the best candidates for learning under conditions of variability. We have shown that SRNs succeed in accounting for the experimental U shape patterns. This is not an easy feat, because SRNs have initial architectural biases toward local dependencies (Chater & Conkey, 1992; Christiansen & Chater, 1999) and because better predictions in SRNs tend to converge towards the optimal conditional probabilities of observing a particular successor to the sequence presented up to that point. This means that minima are located at points in weight space where the activations equal the optimal conditional probability. In fact, activations of output units corresponding to the three final items to be predicted in set-size 2, 6, and 12 settle around.33, which is the optimal conditional probability for $(B|X)$ across conditions. However, n-gram transitional probabilities fail to account for non-adjacent constraints, yielding suboptimal solutions. The networks' ability to predict non-adjacencies is modulated by variability of the intervening element, under conditions of either nil or high variability, achieved by developing separate graded representations in the hidden units. An analysis of hidden unit trajectories over training and control simulations with new embedded elements presented at test suggests that the networks' success at the endpoints of the U curve might be supported by a similar type of learning, thus ruling out a simplistic rote learning explanation for Set-size 1.

We presented a connectionist model that can capture in a single representation both local and non-local properties of the input in a superimposed fashion. This permits it to discover structured sequential input in an implicit, associative way. Together, the experimental and simulation data on the U-curve challenge previous AGL accounts based on one default source of learning. The major implication of this work is that, rather than ruling out associative mechanisms across the board, some statistical learning based on distributional information can account for apparently puzzling aspects of human learning of non-adjacent dependencies. Furthermore, to the extent that these models fit the human data without explicit knowledge, they provide a proof of concept that explicit conscious knowledge may not be necessary to acquire long-distance relations.

Acknowledgments

This work was supported by European Commission Grant HPRN-CT-1999-00065, an institutional grant from the Université Libre de Bruxelles, a Human Frontiers Science Program Grant (RGP0177/2001-B), and Nanyang Technological University's Start-Up-Fund #M4081274. Axel Cleeremans is a Senior Research Associate of the National Fund for Scientific Research (Belgium).

References

Allen, J., & Seidenberg, M.S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *The emergence of language* (pp. 115–151). Mahwah, NJ: Lawrence Erlbaum Associates.

Botvinick, M., & Plaut, D.C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*, 201–233. DOI: 10.1037/0033-295X.113.2.201

Botvinick, M., & Plaut, D.C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395–429. DOI: 10.1037/0033-295X.111.2.395

Brakel, P., & Frank, S.L. (2009). Strong systematicity in sentence processing by simple recurrent networks. In N.A. Taatgen, H. van Rijn, J. Nerbonne & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1599–1604). Austin, TX: Cognitive Science Society.

Chater, N., & Conkey, P. (1992). Finding linguistic structure with recurrent neural networks. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 402–407). Hillsdale, New Jersey: Psychology Press.

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, *35*(1), 26–58.

Christiansen, M.H., Allen, J., & Seidenberg, M.S. (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., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205. DOI: 10.1207/s15516709cog2302_2

Christiansen, M.H., Conway, C.M., & Curtin, S. (2000). A connectionist single-mechanism account of rule-like behavior in infancy. In L. R. Gleitman & A.K. Joshi (Eds.), *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 83–88). Philadelphia, PA: University of Pennsylvania.

Christiansen, M.H., & MacDonald, M.C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning, 59*(Suppl. 1), 126–161. DOI: 10.1111/j.1467-9922.2009.00538.x

Cleeremans, A., Servan-Schreiber, D., & McClelland, J.L. (1989). Finite state automata and simple recurrent networks. *Neural Computation, 1*, 372–381. DOI: 10.1162/neco.1989.1.3.372

Destrebecqz, A., & Cleeremans, A. (2003). Temporal factors in sequence learning. In Luis Jiménez (Ed.), *Attention and implicit learning.* Amsterdam: John Benjamins. DOI: 10.1075/aicr.48.11des

Cottrell, G.W., & Plunkett, K. (1995). Acquiring the mapping from meanings to sounds. *Connection Science*, *6*, 379–412. DOI: 10.1080/09540099408915731

Dell, G.S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*, 149–195. DOI: 10.1207/s15516709cog1702_1

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, *23*, 53–82. DOI: 10.1207/s15516709cog2301_3

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

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

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

Estes, K., Evans, J., Alibali, M., & Saffran, J. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3), 254. DOI: 10.1111/j.1467-9280.2007.01885.x

Farkaš, I., & Crocker, M.W. (2008). Syntactic systematicity in sentence processing with a recurrent self-organizing network. *Neurocomputing*, *71*(7), 1172–1179. DOI: 10.1016/j.neucom.2007.11.025

Frank, M.C., Goldwater, S., Griffiths, T.L., & Tenenbaum, J.B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. DOI: 10.1016/j.cognition.2010.07.005

Frank, S.L. (in press). Getting real about systematicity. In P. Calvo & J. Symons (Eds.), *Systematicity and cognitive architecture: Conceptual and empirical issues 25 years after Fodor & Pylyshyn's challenge to connectionism*. Cambridge, MA: The MIT Press.

Frinken, V., Fischer, A., Manmatha, R., & Bunke, H. (2012). A novel word spotting method based on recurrent neural networks. *IEEE Transactions on, Pattern Analysis and Machine Intelligence*, *34*(2), 211–224. DOI: 10.1109/TPAMI.2011.113

Gaskell, M.G., Hare, M., & Marslen-Wilson, W.D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science*, *19*, 407–439. DOI: 10.1207/s15516709cog1904_1

Gibson, F.P., Fichman, M., & Plaut, D.C. (1997). Learning in dynamic decision tasks: Computational model and empirical evidence. *Organizational Behavior and Human Decision Processes*, *71*, 1–35. DOI: 10.1006/obhd.1997.2712

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431–436. DOI: 10.1111/1467-9280.00476

Harm, M.W., & Seidenberg, M.S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528. DOI: 10.1037/0033-295X.106.3.491

Hinoshita, W., Arie, H., Tani, J., Okuno, H.G., & Ogata, T. (2011). Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network. *Neural Networks*, *24*(4), 311–320. DOI: 10.1016/j.neunet.2010.12.006

Johnstone, T. & Shanks, D.R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology*, *42*, 61–112. DOI: 10.1006/cogp.2000.0743

Jordan, M.I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kinder, A. & Shanks, D.R. (2001). Amnesia and the declarative/procedural distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience*, *13*, 648–669. DOI: 10.1162/089892901750363217

Kirov, C., & Frank, R. (2012). Processing of nested and cross-serial dependencies: An automaton perspective on SRN behaviour. *Connection Science*, *24*(1), 1–24. DOI: 10.1080/09540091.2011.641939

Lashley, K.S. (1951). The problem of serial order in behavior. In L.A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–146). New York, NY: Wiley.

Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York, NY: Wiley.

Maraqa, M., Al-Zboun, F., Dhyabat, M., & Zitar, R.A. (2012). Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. *Journal of Intelligent Learning Systems and Applications*, *4*(1), 41–52. DOI: 10.4236/jilsa.2012.41004

Maskara, A., & Noetzel, A. (1992). Forced simple recurrent neural network and grammatical inference. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 420–425). Hillsdale, NJ: Lawrence Erlbaum Associates.

Miikkulainen, R., & Mayberry III, M. R. (1999). Disambiguation and grammar as emergent soft constraints. In B. MacWhinney (Ed.), *Emergence of language*, 153–176. Mahwah, NJ: Lawrence Erlbaum Associates.

Misyak, J.B., & Christiansen, M.H. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62,* 302–331. DOI: 10.1111/j.1467-9922.2010.00626.x

Misyak, J.B., Christiansen, M.H. & Tomblin, J.B. (2010a). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, Sept.14. DOI: 10.3389/fpsyg.2010.00031.

Misyak, J.B., Christiansen, M.H. & Tomblin, J.B. (2010b). Sequential expectations: The role of prediction- based learning in language. *Topics in Cognitive Science, 2,* 138–153. DOI: 10.1111/j.1756-8765.2009.01072.x

Moss, H.E., Hare, M.L., Day, P., & Tyler, L.K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, *6*, 413–427. DOI: 10.1080/09540099408915732

Munakata, Y., McClelland, J.L., & Siegler, R.S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*, 686–713. DOI: 10.1037/0033-295X.104.4.686

Onnis, L., Christiansen, M.H., Chater, N., & Gómez, R. (submitted). Statistical learning of nonadjacent relations. Submitted manuscript.

Onnis, L., Christiansen, M.H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Preliminary evidence from Artificial Grammar Learning. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society.* Boston, MA: Cognitive Science Society.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 1047–1052). Mahwah, NJ: Lawrence Erlbaum.

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

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*(5), 233–238. DOI: 10.1016/j.tics.2006.03.006

Plaut, D.C., & Kello, C.T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language* (pp. 381–415). Mahwah, NJ: Lawrence Erlbaum Associates.

Redington, M., & Chater, N. (2002). Knowledge representation and transfer in artificial grammar learning (AGL). In R.M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, philosophical, and computational consensus in the making*. Hove: Psychology Press.

Rohde, D.L.T., & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109. DOI: 10.1016/S0010-0277(99)00031-1

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

Saffran, J. (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

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

Si, Y., Xu, J., Zhang, Z., Pan, J., & Yan, Y. (2012). An improved Mandarin voice input system using recurrent neural network language model. In *Computational Intelligence and Security* (*CIS*), *Eighth International Conference on* (pp. 242–246). IEEE.

Socher, R., Manning, C.D., & Ng, A.Y. (2010). Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*. Hilton: Cheakmus.

Sutskever, I., Martens, J., & Hinton, G. (2011). Generating text with recurrent neural networks. In *Proceedings of the 2011 International Conference on Machine Learning* (ICML-2011).

Tabor, W. (2011). Recursion and recursion-like structure in ensembles of neural elements. In H. Sayama, A. Minai, D. Braha, & Y. Bar-Yam (Eds.), *Unifying themes in complex systems. Proceedings of the VIII International Conference on Complex Systems* (pp. 1494–1508). Berlin: Springer.

Takac, M., Benuskova, L, & Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. *Cognition*, *125*, 288–308. DOI: 10.1016/j.cognition.2012.06.006

Vokey, J.R., & Brooks, L.R. (1992). Salience of item knowledge in learning artificial grammar. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 328–344. DOI: 10.1037/0278-7393.18.2.328

Appendix A. Measures of associative learning

Global associative chunk strength (GCS, Knowlton & Squire, 1994) averages the frequencies of all bigrams and trigrams that appear in strings. For instance, one can calculate the GCS for grammatical test items in set-size 2. The form of each test item is $A_iX_jB_i$, with 3 $A_i\_B_i$ dependencies and 2 X_j-elements. A specific item, for instance $A_1X_2B_1$, is composed of 2 bigrams, A_1X_2 and X_2B_1, each repeated 72 times during training, and one trigram $A_1X_2B_1$, repeated 72 times. The GCS measure for this item is obtained by averaging the summed frequencies of each n-gram by the number of n-grams:

$$\frac{freq(A_1X_2)+freq(X_2B_1)+freq(A_1X_2B_1)}{3}=\frac{72+72+72}{3}=72$$

Likewise, the GCS for an ungrammatical test item in set-size 2, say $A_1X_2B_2$, is calculated as follows:

$$\frac{freq(A_1X_2)+freq(X_2B_2)+freq(A_1X_2B_2)}{3}=\frac{72+72+0}{3}=48$$

The Anchor Associative Chunk strength measure (ACS, Reber & Allen, 1978) is similar to the Global Chunk Strength measure, but gives greater weight to the salient initial and final symbols of each string. It is computed by averaging the frequencies of the first and last bigrams and trigrams in each string. In this particular case, because the strings only contain three items the ACS scores are the same as the GCS scores (see Table 2). The first two rows in Table 2 show that GCS/ACS values are always higher for grammatical than for ungrammatical sentences (with a constant ratio of 1.5) and that both values decrease as a function of set-size. Such measures predict that if learners were relying on chunk strength association, their performance should decrease as set-size increases, and thus they do not capture the U shape.

The Novelty measure counts the number of fragments that are new in a sentence presented at test (Redington & Chater, 1996; 2002). This score is 0 for grammatical test strings across conditions, because they do not contain novel fragments and 1 for ungrammatical test strings because they contain one new trigram A_iXB_j. This measure predicts a preference for grammatical strings across conditions, and thus does not capture the U shape either. Yet another measure is novel fragment position (NFP, Johnstone & Shanks, 2001), which counts the number of known fragments in novel absolute position. This score is 0 for both grammatical and ungrammatical test strings, since no fragment appears in a new position with respect to training items and thus cannot account for any differences in grammaticality judgments across conditions. Lastly, Global Similarity (GS) measures the number of letters in a test string that differ from the nearest training string (Vokey & Brooks, 1992). For grammatical test strings this score is 0, and for ungrammatical test items it is 1. Since this value is the same across conditions, GS predicts preference for grammatical strings in all conditions, and again fails to capture the U shape.