Statistical Learning of Nonadjacencies Predicts On-line Processing of Long-Distance Dependencies in Natural Language

Jennifer B. Misyak (jbm36@cornell.edu) and Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853 USA

J. Bruce Tomblin (j-tomblin@uiowa.edu)
Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242 USA

Abstract

Statistical learning (SL) research aims to clarify the potential role that associative-based learning mechanisms may play in language. Understanding learners' processing of nonadjacent statistical structure is vital to this enterprise, since language requires the rapid tracking and integration of long-distance dependencies. This paper builds upon existing nonadjacent SL work by introducing a novel paradigm for studying SL on-line. By capturing the temporal dynamics of the learning process, the new paradigm affords insights into the time course of learning and the nature of individual differences. Across 3 interrelated experiments, the paradigm and results thereof are used to bridge knowledge of the empirical relation between SL and language within the context of nonadjacency learning. *Experiment 1* therefore charts the micro-level trajectory of nonadjacency learning and provides an index of individual differences in the new task. Substantial differences are further shown to predict participants' sentence processing of complex, long-distance natural dependencies in *Experiment 2*. SRN simulations in *Experiment 3* then closely capture key patterns of human nonadjacency processing, attesting to the efficacy of associative-based learning mechanisms that appear foundational to performance in the new, language-linked task.

Keywords: Nonadjacent Dependencies; Sentence Processing; Serial Reaction Time Task; Simple Recurrent Network (SRN)

Introduction

Statistical learning is an inextricably temporal phenomenon, involving the encoding of sequential regularities unfolding over time and space, and the simultaneous shaping of distributional knowledge through ongoing learning experience. Within the past decade, statistical learning (SL) has especially emerged as a key proposed mechanism for acquiring probabilistic dependencies inherent in the time-dependent signal of the speech stream (for reviews, see Gómez & Gerken, 2000; Saffran, 2003).

While traditional artificial grammar learning (AGL; Reber, 1967) tasks have been fruitfully deployed towards studying SL, they fail to provide a clear window onto the temporal dynamics of the learning process. In contrast, serial reaction time (SRT; Nissen & Bullemer, 1987) tasks widely used in sequence-learning research trace individuals' trial-by-trial progress, yet aim to investigate learning for primarily repeating structure. Rarely have methodological advantages of each paradigm been jointly subsumed under a single task for exploring properties of SL.

Nonetheless, notable exceptions include the work of Cleeremans and McClelland (1991), who implemented a noisy finite-state grammar within a visual SRT task to study the encoding of contingencies varying in temporal distance; and of Hunt and Aslin (2001), who employed a visual SRT paradigm for examining learners' processing of sequential transitions varying in conditional and joint probabilities. Howard, Howard, Dennis and Kelly (2008) also adapted the visual SRT to manipulate the types of statistics governing triplet structures; and Remillard (2008) controlled nth-order adjacent and nonadjacent conditional information to probe SRT learning for visuospatial targets. Across these studies, participants evinced complex, procedural knowledge of the sequence-embedded relations upon extensive training over 20, 48, 6 or 4 sessions, respectively, spanning separate days. Reaction time measures throughout exposure enabled insights into the processing of the structural dependencies.

In similar vein, we introduce a new paradigm that directly instantiates an *artificial language* within an adapted SRT task. Distinct from the aforementioned work, the paradigm specifically endeavors to capture the continuous timecourse of statistical processing, rather than contrasting/altering the forms of statistical information. The paradigm is designed for the *briefer* exposure periods prototypic of many AGL studies and flexibly accommodates the use of *linguistic* stimuli-tokens and *auditory* cues. More generally, the task shares similarities to standard AGL designs in the language-like nature of string-sequences, the smaller number of training exemplars, and the greater transparency to natural language structure. Crucially however, it uses the dependent variable of reaction times and a modified two-choice SRT layout to indirectly assess learning while focusing attention through a cover task. By coupling strengths intrinsic to AGL and SRT methods respectively, the 'AGL-SRT paradigm' is intended to complement existing approaches in SL research.

Understanding how learners process nonadjacent relations constitutes an ongoing area of SL work, with importance for theories implicating SL in language acquisition/processing. Natural language abounds with long-distance dependencies that proficient learners must track on-line (e.g., as in subject-verb agreement, clausal embeddings, and relationships between auxiliaries and inflected morphemes). Even with the growing bulk of SL work aiming to address the acquisition of nonadjacencies (e.g., Gómez, 2002; Newport & Aslin, 2004; Onnis, Christiansen, Chater & Gómez, 2003; Pacton & Perruchet, 2008; inter alia), it is yet unknown exactly how such learning unfolds, the precise mechanisms subserving it, and the degree to which SL of

nonadjacencies empirically relates to natural language processing. Our AGL-SRT paradigm offers a novel entry point into this study by augmenting current knowledge with finer-grained, temporal data of how nonadjacency-pairs may be processed over time. As such, Experiment 1 implements Gómez's (2002) high-variability language within the AGL-SRT task to reveal the timecourse of nonadjacent SL. Experiments 2 and 3 then probe the task's relevance to language and its computational underpinnings.

Experiment 1: Statistical Learning of Nonadjacencies in the AGL-SRT Paradigm

In infants and adults, it has been established that relatively high variability in the set-size from which an 'intervening' middle element of a string is drawn facilitates learning of the nonadjacent relationship between the two flanking elements (Gómez, 2002). In other words, when exposed to artificial, auditory strings of the form aXd and bXe, individuals display sensitivity to the nonadjacencies (i.e., the a_d and b_e relations) when elements composing the X are drawn from a large set distributed across many exemplars (e.g., when $|X| = 18$ or 24). Performance is at chance, however, when variability of the set-size for the X is intermediate (e.g., $|X| = 12$) or low (e.g., $|X| = 2$). Similar findings of facilitation from high-variability conditions have also been documented for adults when the grammar is alternatively instantiated with visual shapes as elements (Onnis et al., 2003). Thus, findings have begun to document supportive learning contexts for both infants and adults, but we know little about the timecourse of high-variability nonadjacency learning as it actually unfolds. Here, we address this gap by using the novel AGL-SRT paradigm.

Method

Participants Thirty monolingual, native English speakers from among the Cornell undergraduate population (age: M=20.6, SD=4.2) were recruited for course credit.

Materials During training, participants observed strings belonging to Gómez's (2002) artificial high-variability, nonadjacency language. Strings thus had the form aXd, bXe, and cXf, with initial and final items forming a dependency pair. Beginning and ending stimulus tokens ($a, b, c; d, e, f$) were instantiated by the nonwords *pel, dak, vot, rud, jic,* and *tood;* middle X-tokens were instantiated by 24 disyllabic nonwords: *wadim, kicey, puser, fengle, coomo, loga, gople, taspu, hiftam, deecha, vamey, skiger, benez, gensim, feenam, laeljeen, chila, roosa, plizet, balip, malsig, suleb, nilbo,* and *wiffle.* Assignment of particular tokens (e.g., *pel*) to particular stimulus variables (e.g., the c in cXf) was randomized for each participant to avoid learning biases due to specific sound properties of words. Mono- and bi-syllabic nonwords were recorded with equal lexical stress from a female native English speaker and length-edited to 500 and 600 msec respectively. Ungrammatical items were produced by disrupting the nonadjacent relationship with an incorrect final element to produce strings of the form: **aXe, *aXf, *bXd, *bXf, *cXd* and **cXf*. Written forms of nonwords (in

Arial font, all caps) were presented using standard spelling.
Procedure A computer screen was partitioned into a grid consisting of six equal-sized rectangles: the leftmost column contains the beginning items (a, b, c), the center column the middle items ($X_1...X_{24}$), and the rightmost column the ending items (d, e, f). Each trial began by displaying the grid with a written nonword centered in each rectangle, with each column containing a nonword from a correct and an incorrect stimulus string (foils). Positions of the target and foil were randomized and counterbalanced such that each occurred equally often in the upper and lower rectangles. Foils were only drawn from the set of items that can legally occur in a given column (beginning, middle, end). E.g., for the string *pel wadim rud* the leftmost column might contain PEL and the foil DAK, the center column WADIM and the foil FENGLE, and the rightmost column RUD and the foil TOOD, as shown in Figure 1 across three time steps.
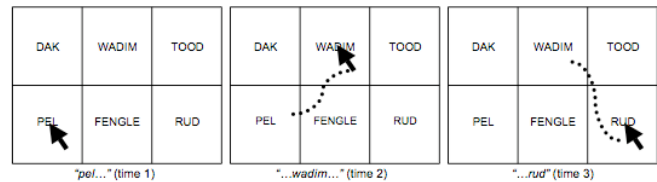

Figure 1: The sequence of mouse clicks associated with a single trial for the auditory stimulus string "*pel wadim rud*".

After 250 msec. of familiarization to the six visually presented nonwords, the auditory stimuli were played over headphones. Participants were instructed to use a computer mouse to click upon the rectangle with the correct (target) nonword as soon as they heard it, with an emphasis on both speed and accuracy. Thus, when listening to *pel wadim rud* the participant should first click PEL upon hearing *pel* (Fig. 1, left), then WADIM when hearing *wadim* (Fig. 1, center), and finally RUD after hearing *rud* (Fig. 1, right). After the rightmost target has been clicked, the screen clears, and a new set of nonwords appears after 750 msec. An advantage of this design is that every nonword occurs equally often (within a column) as target and as foil. This means that for the first two responses in each trial (leftmost and center columns), participants cannot anticipate beforehand which is the target and which is the foil. Following the rationale of standard SRT experiments, however, if participants learn the nonadjacent dependencies inherent in the stimulus strings, then they should become increasingly faster at responding to the final target. The dependent measure is thus the reaction time (RT) for the predictive, final element on each trial, subtracted from the RT for the nonpredictive, initial element to serve as a baseline and control for practice effects.

Each training block involved the random presentation of 72 unique strings (24 strings x 3 dependency-pairs). After exposure to these 432 strings (across the first 6 training blocks), participants were surreptitiously presented with 24 ungrammatical strings, with endings that violated the dependency relations (in the manner noted above). This short ungrammatical block was followed by a final training

('recovery') block with 72 grammatical strings. Block transitions were seamless and unannounced to participants.

Upon completing all 8 blocks, participants were informed that the sequences they heard had been generated according to rules specifying the ordering of nonwords. For an ensuing 'prediction task,' participants were instructed to select string endings for 12 trials upon being cued with only preceding sequence-elements. I.e., participants viewed the same grid display as before and followed the same procedure for the first two string-elements (e.g., *pel wadim…* in Fig. 1) but had to indicate which of the two nonwords in the 3rd column (e.g., TOOD or RUD) they thought best completed the string without hearing the final nonword (and without feedback).

Results and Discussion

Analyses were performed on only accurate string trials (with no more than one selection response for each of the three targets); these comprised grand averages of 90.0% (*SD*=5.6) of training block trials, 84.7% (*SD*=15.7) of ungrammatical trials, and 87.1% (*SD*=12.3) of recovery trials.[1] Mean RT difference scores were then computed for each block.

A one-way repeated-measures analysis of variance (ANOVA) with block as the within-subjects factor was performed. As Mauchly's test indicated a violation of the sphericity assumption (χ^2(27) = 111.82, p <.001), degrees of freedom were corrected using Greenhouse-Geisser estimates (ε = .36). Results indicated that mean RT difference was affected by block, F (2.55, 73.96) = 8.97, p <.001. Figure 2 plots group averages for the mean *RT difference scores* (i.e., initial-element RT minus final-element RT), with positive values reflecting nonadjacency learning. RT differences gradually increased throughout, albeit with an expected decline in the ungrammatical 7th block. Cleeremans and McClelland (1991) have previously found that sensitivity to long-distance contingencies emerges more gradually than for adjacent dependencies; our temporal trajectory in Figure 2 also indicates that sensitivity to nonadjacent dependencies requires considerable exposure (5 blocks on average) before it reliably affects responses.

Planned contrasts confirmed that mean RT differences in the ungrammatical block significantly decreased compared to both the preceding training block, t(29) = 2.11, p =.04, and the following recovery block, t(29) = 3.22, p <.01. Following interpretations in the implicit learning literature for comparing RTs to structured versus unstructured material, this decrement in performance (Block 6 minus Block 7: *M*= -34.8 ms, *SE*=16.5) provides evidence for participants' sensitivity to violations of the sequential structure, with improved performance demonstrated upon the reinstatement of grammatical sequences in the recovery block (Block 8 minus Block 7: *M*= 77.3 ms, *SE*=24.0 ms).

---

[1] As analyzed trials required accuracy for all 3 string-elements composing a string-trial (rather than for single-selection responses defining one 'trial' in standard SRT designs), this criterion is quite conservative, and may underestimate participants' total accuracy across all single responses. E.g., final-element selection accuracy across trial-types was 95.9% (2.4), 93.2% (6.5), and 94.2% (6.1).

Prediction task accuracy scores averaged 61.1% (SD=21.4%) reflecting substantial interindividual variation. Group-level performance was above chance, (t(29) = 2.85, p <.01), providing another gauge of nonadjacency learning. Such scores further provide a sensitive index of individual differences for the on-line language processing of complex long-distance dependencies, as the next experiment shows.
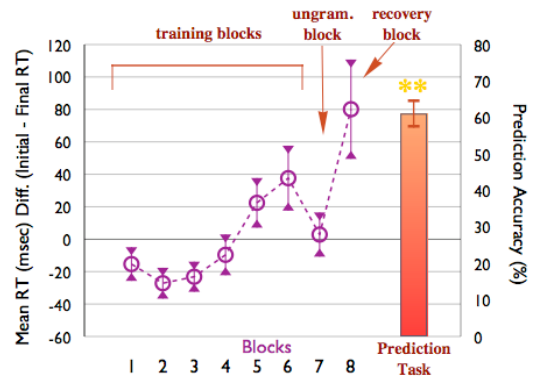

Figure 2: Group learning trajectory (as a plot of mean RT differences) and prediction accuracy in Experiment 1.

Experiment 2: Individual Differences in Language Processing and Statistical Learning

Individual differences in tracking long-distance dependencies in natural language have been extensively studied in relation to the contrastive processing of subject and object relatives. Object relative (OR) sentences (illustrated in *2*) involve a head-noun that is the object of an embedded clause, and are generally more difficult to process and comprehend than subject relatives (SRs; such as *1*), in which the head-noun is the subject of the modifying clause. ORs are of keen interest here because successfully tracking their structure entails integrating nonadjacent dependencies over lexical constituents (i.e., relating the embedded verb to the nonlocal head-noun and relating the head-noun to the main verb from across the embedded clause).

(1) The reporter <u>that attacked the senator</u> admitted the error.
(2) The reporter <u>that the senator attacked</u> admitted the error.

Differential processing difficulty between ORs and SRs is most acute at the main verb, where protracted reading times (RTs) for ORs are evidenced. Individual differences in the degree of comparative difficulty have been first reported by King and Just (1991) and linked to variations in verbal working memory (vWM) as assessed by a reading span task. Interpretations of these findings, however, have been in dispute between experience-based versus capacity-based accounts (e.g., Just & Carpenter, 1992; MacDonald & Christiansen, 2002; see also Waters & Caplan, 1996).

While capacity-based views impute low-span individuals' poorer processing of ORs to limitations in memory resources, experience-based views emphasize experiential learning factors that modulate the processing difficulty that readers encounter. In support of the latter approach, MacDonald and Christiansen (2002) conducted simple

recurrent network (SRN) simulations that reproduced the SR/OR RT patterns of low- and high-span individuals as a function of the amount of training received by the networks. In addition, a human training study by Wells, Christiansen, Race, Acheson and MacDonald (2009) showed that greater SR/OR reading experience (compared to that of a control condition) tuned RT profiles towards resembling those of high-span individuals and qualitatively fit the performance of the aforementioned SRNs after the most training epochs.

These studies suggest that SL plays a crucial underlying role in shaping readers' experience of the distributional constraints that govern the less frequent and irregular ORs, which in turn facilitates subsequent RTs. If SL is indeed an important mechanism for such processing phenomena and is meaningfully captured by the new AGL-SRT task, then individual differences in nonadjacent SL (as observed and indexed in Exp. 1) should systematically contribute towards interindividual variation for the ability to track the nonlocal dependency structure of OR sentences. Exp. 2 thus aims to empirically test the strength of this predicted relationship.

Method

Participants Nineteen of the last 20 participants (age: M=20.0, SD=1.4) in Experiment 1 participated afterwards in this experiment for additional credit. Data from one participant was omitted due to equipment malfunction.

Materials Two experimental sentence lists were prepared, each incorporating 12 initial practice items, 40 experimental items (20 SRs, 20 ORs), and 48 filler items. Yes/No comprehension probes accompanied each sentence item. The SR/OR sentence pairs were taken from Wells et al. (2009) and counterbalanced across the two lists. Semantic plausibility information for subject/object nouns was controlled in the experimental materials.

Procedure Each participant was randomly assigned to a sentence list, whose items were presented in random order using a standard word-by-word, moving-window paradigm for self-paced reading (Just, Carpenter & Woolley, 1982). Millisecond RTs for each sentence-word and accuracy for each following comprehension question were recorded.

Results and Discussion

Raw RTs corresponding to practice items and those in excess of 2500 ms (0.86% of data) were excluded from analyses. RTs were length-adjusted by computing a regression equation per participant based on the character-length of a word, and subtracting observed RT values from predicted values (Ferreira & Clifton, 1986). Means from residual RTs were then calculated for the same sentence regions as used in Wells et al. (2009) and prior related work.

Overall comprehension rate was high (87.3%). Consistent with past studies, comprehension was poorer for ORs (75.8%) compared to SRs (86.1%). To test the involvement of SL in mediating individual differences in corresponding RT patterns, participants were first classified as 'low' or 'high' in SL skill according to their prediction task scores from Exp. 1 (with 50% as the cutoff-level). RTs from 'low

pred' (n=11, M= 42.4%, SD=8.7) and 'high pred' (n=7, M= 73.8%, SD=14.8) participants were then compared.

While the two groups did not differ on their processing of SR regions, RTs considerably diverged at the main verb of ORs, as depicted in Figure 3. This performance contrast for ORs (and lack thereof for SRs) precisely mirrors the reading patterns documented in the literature for those with 'low' and 'high' vWM span scores respectively. Importantly then, individual differences in SL prediction task scores were not predictive of RTs for any SR/OR sentence regions *except*, crucially, at the main verb of ORs (R^2= .34, p= .01)—the anticipated locus of observed processing difficulty.

These findings suggest that skill in learning and applying statistical knowledge of distributional regularities, as indexed by prediction task scores from the novel AGL-SRT paradigm, is substantially involved in natural language processing of relative clauses. This conclusion is also supported by results from an individual-differences study by Misyak and Christiansen (2007), in which both adjacent and nonadjacent statistical learning performance was an even better predictor of sentence comprehension than vWM span scores. The current study thus expands on those findings by documenting that differences in nonadjacent SL vary systematically with the on-line tracking and integration of nonadjacent dependencies exemplified by OR sentences.
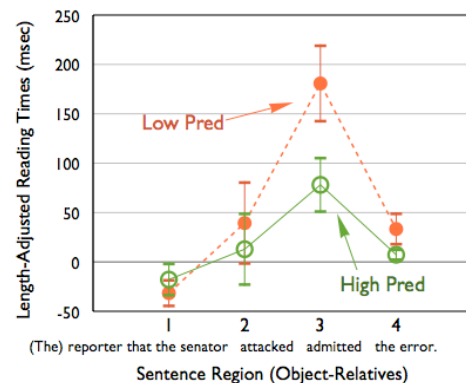

Figure 3: Length-adjusted reading times by sentence region of obj.-relatives for 'low' and 'high' pred score participants.

Experiment 3: Computational Simulations of On-line Nonadjacency Learning

While Experiment 2 supports the relevance of the new AGL-SRT task for the processing of complex long-distance dependencies in natural language, the kind of computational mechanisms underpinning task performance remains to be probed. MacDonald and Christiansen's (2002) simulations of relative clause processing suggest that mechanisms akin to those of simple recurrent networks (SRNs; Elman, 1990) may suffice. Moreover, Cleeremans and McClelland (1991) have formerly shown that the SRN can capture performance on AGL-like SRT tasks. We thus chose to closely model on-line performance from our task with SRN simulations based on the *exact* same exposure and input as in the human case.

The SRN is essentially a standard feed-forward network equipped with context units containing a copy of hidden unit activation at the previous timestep, thus providing partial recurrent access to prior internal states. The context layer's limited maintenance of sequential information over past timesteps allows the SRN to potentially discover temporal contingencies spanning varying distances in the input. Next, we use the SRN's graded output values and prediction-based learning mechanism to model human RTs and prediction scores from Experiment 1.

Method

Networks Simulations were conducted with 30 individual networks, one corresponding to each human participant, and each randomly initialized with a different set of weights within the interval (-1,1) to approximate learner differences. Localist representations were employed for the 30 input and output units, with one unique unit corresponding to each nonword item. The hidden layer had 15 units. The networks were trained using standard backpropagation with a learning rate of 0.1 and momentum at 0.8.

Materials The SRNs received the same input as human participants, presented using the same randomization process as in Experiment 1, and tested on the same 'prediction task' strings (with the same target-foil pairings).

Procedure SRNs were trained on the strings following an identical trial-type sequence as that in Exp. 1 and given a subsequent 'prediction task.' Networks received the exact same amount of exposure to the statistical dependencies as the human participants (i.e., 6 grammatical blocks of 72 string-trials, an ungrammatical block of 24 trials, a recovery block of 72 trials, and a 12-item prediction task)—and no additional training. Context units were reset between string-sequences by setting values to 0.5; this simulated the screen-clear and between-trial pauses that human participants observed. Weight changes were updated continuously throughout training, except for the prediction task items at the very end, when weights were 'frozen' (reflecting the fact that human participants received no auditory input/feedback for selecting the final elements of prediction-task strings).
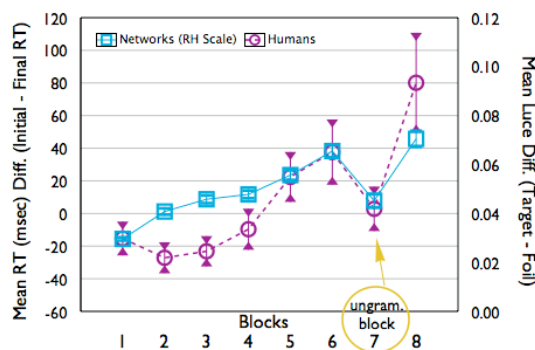

Figure 4: Comparison of group learning trajectories.

Results and Discussion

The networks' continuous outputs were recorded, and performance was evaluated by computing a Luce ratio

difference score for string-final predictions on each trial. A Luce ratio is calculated by dividing a given output-unit's activation value by the sum of the activation values of all output units. During processing, the representation formed at the output layer of the SRN approximates a probability distribution for the network's prediction of the next element. Thus, on the timestep where a middle (*X*) element is received as input, if the network has become sensitive to the nonadjacent dependencies, it should most strongly activate the output unit corresponding to the correct, upcoming string-final nonword. The Luce ratio essentially quantifies the proportion of total activity owned by this output unit.

To approximate human RT difference scores, we subtracted the Luce ratio for the foil unit from the Luce ratio for the target unit. Since networks cannot erroneously select a foil in the same way that humans occasionally do (and which were excluded from analyses, as noted earlier and in line with standard SRT protocol), accurate trials for the networks were defined as those in which the Luce ratio for the target exceeded that for the foil. As in Exp. 1, only responses/outputs from accurate trials were analyzed.

A one-way repeated-measures ANOVA with block as the within-subjects factor was performed. As Mauchly's test indicated a violation of the sphericity assumption ($\chi^2(27) = 66.947$, $p < .001$), degrees of freedom were corrected using Greenhouse-Geisser estimates ($\varepsilon = .60$). There was a main effect of block on mean Luce ratio difference, $F(4.21, 121.96) = 35.57$, $p < .001$. As in the human case, difference scores gradually increased, with a performance decrement in the 7th (ungrammatical) block. This drop was significant in relation to both the preceding and succeeding grammatical blocks, $t(29) = 6.76$, $p < .0001$; $t(29) = 7.80$, $p < .0001$.

As the analog to the human prediction task, in which SRNs received the same test-strings with foil-pairings as the humans, we considered the network's selection to be the nonword corresponding to the unit with a higher Luce ratio (from among the 2 choices for an ending). Prediction task accuracy as a proportion correct out of the 12 items was then computed accordingly. The SRNs' scores averaged 56.4% ($SD=13.4\%$), which was above chance-level, $t(29) = 2.61$, $p = .01$. The networks' score distribution was also not significantly different from that of humans', $t(58) = 1.025$, $p > .30$. Although the networks exhibited somewhat less variability, they captured the identical full range of human performance from 25 - 100% accuracy.

The networks' mean Luce ratio difference scores across blocks are plotted in Figure 4, alongside the human learning trajectory from Exp. 1.[2] Both trajectories are indicative of a gradually developing sensitivity to the nonadjacent dependencies, with a steeper ascent from blocks 4 to 6. The simulated block scores further account for 78% of the variance in human RT difference scores ($p < .01$).

---

[2] Because the learning metric for humans subtracts final- from initial-element RTs (to control for potential motor effects) whereas that for the SRNs uses only final-element values, Y-axes are equalized with block 1 level performance as the baseline.

General Discussion

Nonadjacent dependency learning was investigated here across three interconnected experiments, using results from a novel AGL-SRT paradigm. The new task investigated individuals' learning of nonadjacencies as it unfolded on-line. Task performances were further shown to predict processing for complex, long-distance dependencies occurring in natural language, as well as to compellingly appear to recruit upon the kind of associative-based learning principles exemplified by SRNs.

Our close modeling of human performance with SRNs further argues against the assumption that vWM capacity operates as a basic constraint for results in Exp. 1 and 2; it also establishes a connection with results from MacDonald and Christiansen (2002) in terms of common mechanisms. Their SRN simulations had predicted that ORs should be differentially affected by increased exposure to relative-clause sentences. Wells et al. (2009) empirically confirmed those predictions and further hypothesized that SL may be centrally involved—but did not otherwise speak to what the underlying mechanisms may be. Our Exp. 2, however, directly supports Wells et al.'s hypothesis. Namely, SL prediction performance for high- and low-performing individuals on SR/OR processing closely conformed to the pattern obtained for participants measured to have high/low vWM spans in King and Just (1991), as well as those of the high/low experience manipulations for SRNs and humans in MacDonald and Christiansen and Wells et al., respectively. Together with previous findings that SL overall is a better predictor of sentence processing skills than vWM (Misyak & Christiansen, 2007), these results provide converging evidence for SL as a key contributing factor to individual differences in language processing.

But how do high- and low-SL performers differ? Added inspection of micro-level trajectories from Exp. 1 for high/low SL groups reveals distinct differences during non-adjacency learning. Thus, there are contrasts in the shape of the SL training trajectory, final training performance, and the response to ungrammatical items. In particular, the low-SL performers do not show evidence of learning until the final block, contributing to the strong recovery effect on this block observable in Figure 2. As in this paper, future work studying such *SL differences* (using sensitive paradigms and computational modeling) should be fruitful for further elucidating the interrelationships among SL, language, and nonadjacency processing, as well as the extent of their shared dependence on complex, association-based learning mechanisms (as captured by networks like the SRN).

References

Cleeremans, A. & McClelland, J.L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120*, 235-253.

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

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348-368.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 13, 431-436.

Gómez, R.L. & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4,* 178-186.

Howard, J.H., Jr., Howard, D.V., Dennis, N.A. & Kelly, A.J. (2008). Implicit learning of predictive relationships in three-element visual sequences by young and old adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1139-1157.

Hunt, R.H. & Aslin, R.N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130,* 658-680.

Just, M.A. & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Just, M.A., Carpenter, P.A. & Woolley, J.D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*, 228-238.

King, J. & Just, M.A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30*, 580-602.

MacDonald, M.C. & Christiansen, M.H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review, 109*, 35-54.

Misyak, J.B. & Christiansen, M.H. (2007). Extending statistical learning farther and further: Long-distance dependencies, and individual differences in statistical learning and language. *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1307-1312). Austin, TX: Cognitive Science Society.

Newport, E.L. & Aslin, R.N. (2004). Learning at a distance I. Statistical learning of nonadjacent dependencies. *Cognitive Psychology, 48*, 127-162.

Nissen, M.J. & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*, 1-32.

Onnis, L., Christiansen, M.H., Chater, N. & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial language learning. *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 886-891). Mahwah, NJ: Lawrence Erlbaum Associates.

Pacton, S. & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 80-96.

Reber, A. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.

Remillard, G. (2008). Implicit learning of second-, third-, and fourth-order adjacent and nonadjacent sequential dependencies. *The Quarterly Journal of Experimental Psychology, 61*, 400-424.

Saffran, J.R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12,* 110-114.

Waters, G.S. & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Quarterly Journal of Experimental Psychology, 49*, 51-79.

Wells, J.B., Christiansen, M.H., Race, D.S., Acheson, D.J. & MacDonald, M.C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*, 250-271.