Topics in Cognitive Science 2 (2010) 138–153

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2009.01072.x

Sequential Expectations: The Role of Prediction-Based Learning in Language

Jennifer B. Misyak,[a] Morten H. Christiansen,[a] J. Bruce Tomblin,[b]

[a]*Department of Psychology, Cornell University*
[b]*Department of Communication Sciences and Disorders, University of Iowa*

Received 15 September 2009; received in revised form 26 October 2009; accepted 3 November 2009

Abstract

Prediction-based processes appear to play an important role in language. Few studies, however, have sought to test the relationship within individuals between prediction learning and natural language processing. This paper builds upon existing statistical learning work using a novel paradigm for studying the on-line learning of predictive dependencies. Within this paradigm, a new ''prediction task'' is introduced that provides a sensitive index of individual differences for developing probabilistic sequential expectations. Across three interrelated experiments, the prediction task and results thereof are used to bridge knowledge of the empirical relation between statistical learning and language within the context of nonadjacency processing. We first chart the trajectory for learning nonadjacencies, documenting individual differences in prediction learning. Subsequent simple recurrent network simulations then closely capture human performance patterns in the new paradigm. Finally, individual differences in prediction performances are shown to strongly correlate with participants' sentence processing of complex, long-distance dependencies in natural language.

Keywords: Prediction; Sentence processing; Language comprehension; Statistical learning; Nonadjacent dependencies; Serial reaction time task; Simple recurrent network

1. Introduction

Most individuals can relate to the common, albeit occasionally vexing, experience of having someone else anticipate and finish one's own sentence before one has completed saying it. Such behavior is but one simple reflection of the human ''drive to predict,'' which may serve as a ''powerful engine for learning and provides important clues to latent abstract

Correspondence should be sent to Jennifer B. Misyak, Department of Psychology, B72 Uris Hall, Cornell University, Ithaca, NY 14853. E-mail: jbm36@cornell.edu

structure'' (Elman, 2009, p. 572). The broader processes underlying such ordinary acts have accordingly received attention as an integral component for successful learning, understanding, and use of language. For example, implicit learning of sequential regularities has been linked to an individual's ability to utilize contextual and lexically predictive information in comprehending spoken language; listeners who are better at extracting statistical relationships contained within an aural sequence are also more adept in predicting the sentence-final words of a noisy speech signal (Conway, Bauernschmidt, Huang, & Pisoni, in press). Across other areas of language, empirical data suggest that learned knowledge of probabilistic structure forms the basis for generating implicit expectations of upcoming linguistic input, and that the on-line engagement of such predictive skills comprises an important role in language acquisition and processing (for reviews, see Federmeier, 2007; Kamide, 2008; Van Berkum, 2008).

Statistical learning mechanisms that have been proposed for tracking predictive dependencies in language (Saffran, 2001; for reviews, see Gómez & Gerken, 2000; Saffran, 2003) may thus be viewed as tapping into this prediction-based process. More generally, outside of language, sequence-learning work has similarly examined basic abilities for the rapid anticipation of discrete, temporal elements under incidental learning conditions. However, while traditional artificial grammar learning (AGL; Reber, 1967) tasks have been fruitfully deployed towards studying statistical learning, 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 have been used widely in sequence-learning research to trace individuals' trial-by-trial progress, but primarily with a focus on learning fixed, repeated structure. Despite their natural commonalities then, rarely have methodological advantages of each paradigm been jointly subsumed under a single task for exploring the on-line development of prediction-based learning.

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. Moreover, Howard, Howard, Dennis, and Kelly (2008) adapted the visual SRT task to manipulate the types of statistics governing triplet structures; and Remillard (2008) controlled *n*th-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 collected throughout exposure enabled insights into the processing of the predictive dependencies.

In a similar vein, we employ a novel paradigm that directly implements an *artificial language* within a two-choice SRT task. Distinct from previous statistical learning methods, our paradigm specifically aims to reveal the continuous timecourse of statistical processing, rather than contrasting or altering the types of statistics. The paradigm is designed for the *briefer* exposure periods typical 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 an adapted 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 to research on the statistical learning of predictive relations.

Understanding how learners process nonadjacent dependencies constitutes an ongoing area of such work, with importance for theories implicating statistical learning in language. Natural language characteristically contains many long-distance dependencies that proficient learners need to track on-line (e.g., subject-verb agreement, embedded clauses, and relations between auxiliaries and inflectional morphemes). Even with the growing bulk of statistical learning 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 statistical learning of nonadjacencies empirically relates to natural language processing.

Our AGL-SRT paradigm offers a novel entry point into the study of statistical nonadjacency learning by augmenting current knowledge with finer-grained, temporal data to illuminate how nonadjacent dependencies may be processed and anticipated over time. As such, Experiment 1 studies the timecourse of nonadjacency learning, using our novel AGL-SRT paradigm and incorporating a ''prediction task'' (rather than the kind of standard grammaticality test typically used; e.g., Gómez, 2002). Subsequently, Experiment 2 shows how the prediction-based, associative learning principles exemplified by simple recurrent networks closely accommodate human performances on this prediction task. Experiment 3 then probes the relevance of statistical prediction-task performance to on-line natural language processing.

2. 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). That is, when aurally familiarized to artificial strings of the form *aXd* and *bXe*, individuals show sensitivity to the nonadjacencies (i.e., the *a_d* and *b_e* dependencies) when the set of elements from which *X* is drawn comprise a large set of exemplars (e.g., $|X| = 18$ or 24). Performance is poorer, however, when variability of the set-size for the *X* is intermediate (e.g., $|X| = 12$) or low (e.g., $|X| = 2$). Similar facilitation in 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, although past research has begun to document learning in specific contexts for both infants and adults, we know little about the timecourse for

acquiring predictive nonadjaencies as it actually unfolds. Here, we employ our novel AGL-SRT paradigm towards that aim.

2.1. Method

2.1.1. Participants

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

2.1.2. Materials

Throughout training, participants observed auditory-visual strings (composed of three nonwords) belonging to the artificial high-variability, nonadjacency language of Gómez's (2002). Strings therefore had the form *aXd*, *bXe*, and *cXf*, with ending nonword-items (*d*, *e*, *f*) predictably dependent upon beginning nonword-items (*a*, *b*, *c*). Monosyllabic nonwords (*pel*, *dak*, *vot*, *rud*, *jic*, and *tood*) instantiated the string-initial and final stimulus tokens (*a*, *b*, *c*; *d*, *e*, *f*); bisyllabic 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*) instantiated the set of 24 middle *X*-tokens. The assignment of particular tokens (e.g., *pel*) to specific stimulus variables (e.g., the *c* in *cXf*) was randomized across participants to avoid learning biases attributable to the specific sound properties of words. Auditory forms of the nonwords were recorded by a female native English speaker with equal lexical stress and length-edited to 500 and 600 ms for mono- and bi-syllabic nonwords, respectively. Written forms of nonwords were presented in Arial font (all caps) with standard spelling and appeared on a computer screen that was partitioned into a 2×3 grid of uniform rectangles, as depicted in Fig. 1. The leftmost column of the computer grid contained only the initial items of strings (*a*, *b*, *c*), the center column the middle items

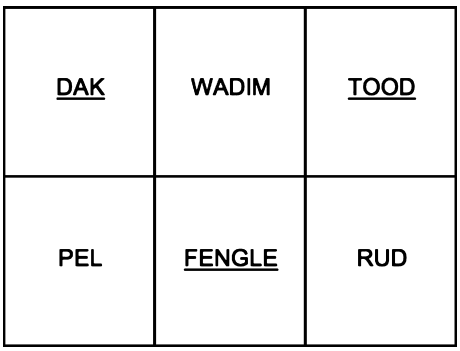

Fig. 1. The grid display for presenting the stimulus strings on each trial. In this example, ''DAK'' and ''PEL'' are initial-string items (*a, b,* or *c* elements) appearing in the leftmost column; ''FENGLE'' and ''WADIM'' are middle-string items (belonging to the set of 24 *X*-elements) appearing in the center column; and ''TOOD'' and ''RUD'' are final-string items (*d, e,* or *f* elements) appearing in the rightmost column. For expository purposes only, some nonwords are underlined here to distinguish the target string (*dak fengle tood*) from the foil string (*pel wadim rud*) in this example.

($X_1 \ldots X_{24}$), and the rightmost column the final items (*d*, *e*, *f*). Ungrammatical strings were generated by substituting an incorrect final element that disrupted the nonadjacency relationship, thus producing strings of the form: **aXe*, **aXf*, **bXd*, **bXf*, **cXd*, and **cXf*.

2.1.3. Procedure

Each trial began by displaying the computer grid with a written nonword centered in each rectangle, with each column containing a nonword from a correct (target) and an incorrect (foil) stimulus string. Positions of targets and foils were randomized and counterbalanced such that they were contained equally often within the upper and lower rectangles. Only the set of items that could legally occur within a given column (initial, middle, final) were used to draw the foils. For example, for the string *dak fengle tood*, the leftmost column might display DAK and the foil PEL, the center column FENGLE and the foil WADIM, and the rightmost column TOOD and the foil RUD, as shown in Fig. 1.

After 250 ms of familiarization to the six written nonwords, auditory versions of the three nonwords were played over headphones. Participants used a computer mouse to click inside the rectangle containing the correct (target) written nonword as soon as they heard it, with instructions emphasizing both speed and accuracy. The first nonword (e.g., *dak*) was played automatically after the familiarization period, whereas the subsequent two nonwords were played once the participant had responded to the previously played word (e.g., *fengle* was played after a response was recorded for *dak*, and *tood*, in turn, after the participant responded to *fengle*). Thus, when listening to *dak fengle tood*, the participant should first click DAK upon hearing *dak* (Fig. 2, left), then FENGLE when hearing *fengle* (Fig. 2, center), and finally TOOD after hearing *tood* (Fig. 2, right). After the participant clicks the rightmost target, the screen clears and a new set of nonwords appears 750 ms later.

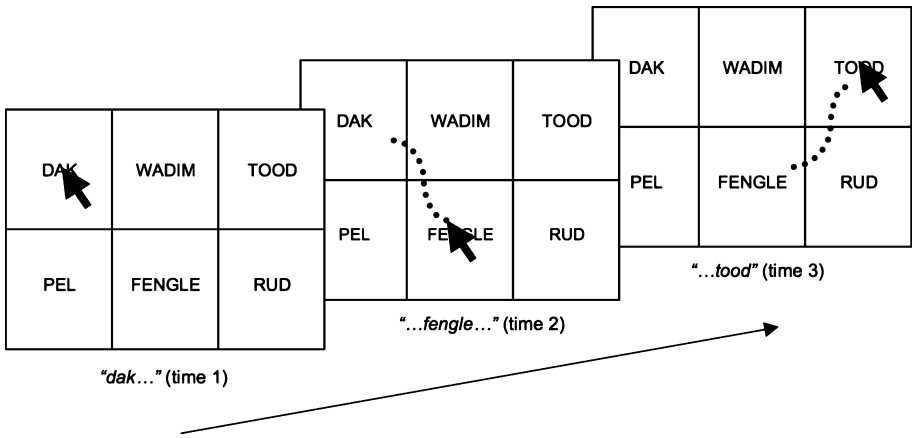

Fig. 2. The sequence of mouse clicks associated with the auditory stimulus string ''*dak fengle tood*'' for a single trial. All trials for each of the blocks (training, ungrammatical, and recovery) followed this general pattern of sequence clicks (from left, to center, to right column clicks corresponding to the selections for the respective elements of a target string).

Per design, each nonword occurs equally often (within a column) as a target and as a foil. Thus, participants cannot anticipate beforehand which is the target and which is the foil for the first two responses of a given trial (leftmost and center columns). However, following the rationale of standard SRT experiments, if participants learn to anticipate the nonadjacent dependencies inherent in the stimulus strings, then they should respond increasingly faster to the final target. As our dependent measure, we thus recorded on each trial the reaction time (RT) for the predictive, final element, subtracted from the RT for the nonpredictive, initial element to control for practice effects and serve as a baseline.

Participants were first exposed to six training blocks, each of which consisted of a random presentation of 72 unique strings (24 strings × 3 dependency-pairs), for exposure to a total of 432 grammatical strings. After this, participants were presented with 24 ungrammatical strings, with endings that violated the nonadjacent dependency (in the manner noted above). A final training ''recovery'' block of 72 grammatical strings then followed this brief ungrammatical block. Transitions between all blocks were seamless and unannounced.

Upon completing the eight exposure blocks, participants performed the ''prediction task'' of key interest here because it provides a direct measure of the degree to which participants have learned the nonadjacency patterns. They were told that there were rules specifying the ordering of nonwords for the auditory sequences, and were asked to indicate the endings for 12 subsequent strings upon being cued with only the first two sequence-elements. In other words, participants observed the same grid display as before and followed the same procedure for the nonpredictive initial and middle columns (e.g., selections corresponding to *dak fengle…* in Fig. 2), but then they had to select which nonword in the predictive final column (e.g., TOOD or RUD) they thought best completed the string without hearing the ending (and without feedback).

2.2. Results

Since instructions emphasized speed in addition to accuracy, there was a small proportion of errors made by participants, as is commonly reported in SRT studies. Thus, only accurate string trials (with only one selection response for each of the three targets) were used for analyses. These averaged 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] Final-element RTs were subtracted from initial-element RTs on each trial, with means of these resultant RT difference scores computed for each block. Fig. 3 plots group averages for these difference scores, with positive values reflecting nonadjacency learning.

A one-way repeated-measures analysis of variance (ANOVA) with block as the within-subjects factor was performed on mean RT difference scores. Mauchly's test indicated a violation of the sphericity assumption ($\chi^2(27) = 111.82$, $p < 0.001$), so Greenhouse-Geisser estimates ($\varepsilon = 0.36$) were used to correct degrees of freedom. There was a significant effect of block on RT scores, $F(2.55, 73.96) = 8.97$, $p < 0.001$. As shown in Fig. 3, RT differences gradually increased across blocks, albeit with an expected performance decrement in the ungrammatical seventh block. As also evidenced by the group trajectory, sensitivity to nonadjacent dependencies required considerable exposure (an average of five blocks) before

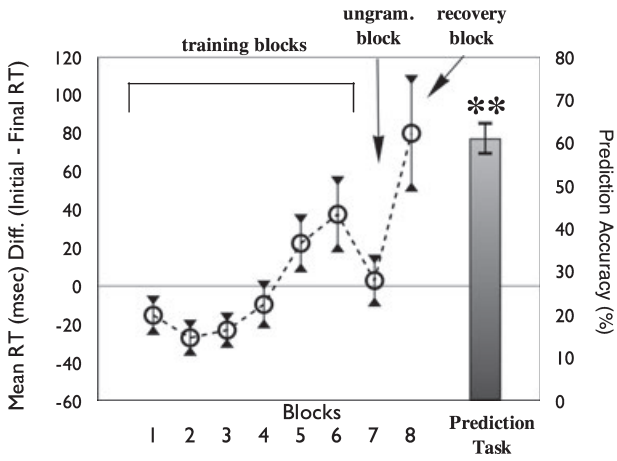

Fig. 3. Group learning trajectory (as a plot of mean RT difference scores) and prediction accuracy in Experiment 1.

reliably affecting responses; this is consistent with Cleeremans and McClelland's (1991) finding that learning for long-distance contingencies emerges less rapidly than for adjacent dependencies.

Following interpretations in the sequence learning literature for comparing RTs to structured versus unstructured material (e.g., Thomas & Nelson, 2001), we specifically assessed performance differences across the final training block, ungrammatical block, and recovery block. Planned contrasts confirmed that mean RT differences significantly decreased in the ungrammatical block compared to performances in both the preceding training block, $t(29) = 2.11$, $p = 0.04$, and succeeding recovery block, $t(29) = 3.22$, $p < 0.01$. This relative performance drop in the ungrammatical block (Block 6 minus Block 7: $M = -34.8$ ms, SE = 16.5; Block 8 minus Block 7: $M = 77.3$ ms, SE = 24.0 ms) provides a confirmation of nonadjacency learning using an established SRT measure.

Of central focus to the interrelated experiments that follow next, accuracy scores on the prediction task were calculated for each individual. Participants averaged 61.1%, with a large standard deviation (21.4%) and group range (25–100%) reflecting substantial interindividual variation. Group-level performance was above chance [$t(29) = 2.85$, $p < 0.01$] providing a gauge of predictive skills for anticipating the statistical nonadjacencies. But what kind of computational mechanism may subserve the kind of learning evidenced by this prediction task and, more generally, by the on-line AGL-SRT task? We address this question in Experiment 2, before going on to show in Experiment 3 that the performance on the prediction task provides a sensitive index of individual differences in on-line language processing.

3. Experiment 2: Computational simulations of on-line nonadjacency learning

The new paradigm in Experiment 1 highlights the gradual statistical learning of nonadjacencies in prediction-based performance; however, the computational mechanisms

that can accommodate such findings remain to be investigated. Cleeremans and McClelland (1991) have previously shown that the simple recurrent network (SRN; Elman, 1990) can capture performance on AGL-like SRT tasks. Furthermore, the anticipation of unfolding temporal structure and implicit prediction-based feedback are distinctive, fundamental features of the SRN's associative architecture (see, e.g., the discussion in Altmann & Mirković, 2009). 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.

3.1. Method

3.1.1. 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. 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.

3.1.2. 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).

3.1.3. Procedure

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 carried out 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).

3.2. Results

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, 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 Experiment 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 ($\chi^2(27) = 66.947$, $p < 0.001$), degrees of freedom were corrected using Greenhouse-Geisser estimates ($\varepsilon = 0.60$). There was a main effect of block on mean Luce ratio difference, $F(4.21,\ 121.96) = 35.57$, $p < 0.001$. As in the human case, difference scores gradually increased, with a performance decrement in the seventh (ungrammatical) block. This drop was significant in relation to both the preceding and succeeding grammatical blocks, $t(29) = 6.76$, $p < 0.0001$; $t(29) = 7.80$, $p < 0.0001$.

The networks' mean Luce ratio difference scores across blocks are plotted in Fig. 4, alongside the human learning trajectory from Experiment 1.[2] Both trajectories are indicative of a gradually developing sensitivity to the nonadjacent dependencies, with a steeper ascent

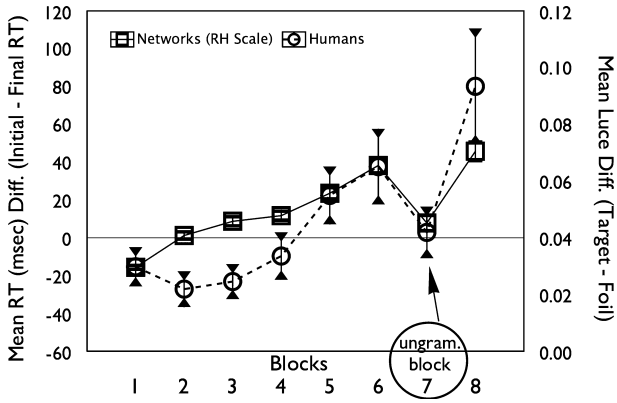

Fig. 4. Comparison of group learning trajectories for SRN (squares) and human (circles) learners.

from blocks 4 to 6. The simulated block scores further account for 78% of the variance in human RT difference scores ($p < 0.01$).

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 two 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, $t(29) = 2.61$, $p = 0.01$. As seen in Fig. 5, the distribution of the networks' prediction scores were also not significantly different from that of humans', $t(58) = 1.025$, $p > 0.30$. Although the networks exhibited somewhat less variability, they captured the identical full range of human performance from 25% to 100% accuracy. Thus, the SRN is able to closely match human performance both across training in the AGL-SRT task as well as on the prediction task. Given that this type of connectionist model has been used extensively to model the processing of nonlocal dependencies in natural language (e.g., Christiansen & Chater, 1999; Christiansen & MacDonald, 2009; Elman, 1991; MacDonald & Christiansen, 2002), we next explore whether the ability to predict correct nonadjacency relations in Experiment 1 is associated with the processing of long-distance dependencies in language.

4. Experiment 3: Individual differences in language processing and statistical learning

While Experiment 2 attests to the kind of computational mechanisms that may subserve performance on the AGL-SRT and prediction tasks, the relevance of the new paradigm for the processing of complex long-distance dependencies in natural language remains to be probed. In the language literature, individual differences have been prominently studied within the context of subject-object relative (OR) clause processing phenomena. Center-embedded OR sentences (illustrated in 2) are generally more difficult to process and comprehend than subject relative sentences (SRs; such as 1), with the structural difference

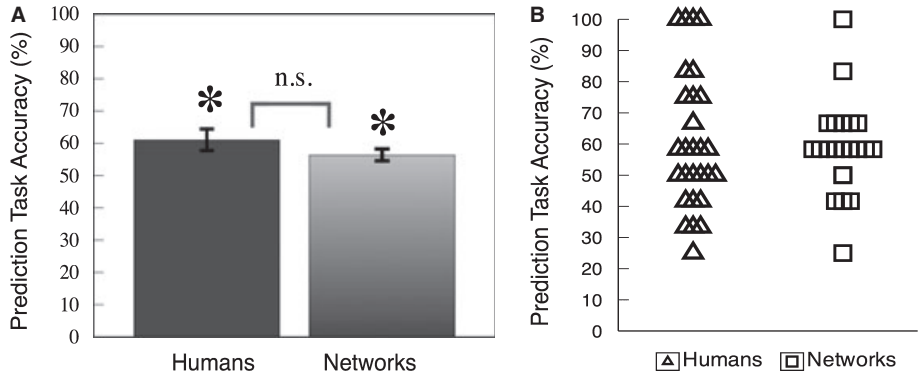

Fig. 5. Prediction task means for humans and networks (A) and corresponding score distributions of both groups (B).

between the two residing in how the embedded verb (*attacked*) relates to its object (but see Reali & Christiansen, 2007). For ORs, the embedded verb enters into a nonadjacent dependency with the nonlocal head-noun (*reporter*), whereas for SRs the embedded verb's object (*senator*) is situated more locally. The greater processing difficulty associated with ORs can be construed as a reflection of changing, probabilistic expectations for the continuation of the sentence as its temporarily indeterminate (and relative to SRs, less frequent and irregular) structure unfolds (Gennari & MacDonald, 2008).

1. The reporter *that attacked the senator* admitted the error.
2. The reporter *that the senator attacked* admitted the error.

The locus of this greater processing difficulty for ORs compared to SRs is evidenced at the main verb, where reading times (RTs) for ORs are protracted. King and Just (1991) first reported individual differences in the degree of comparative difficulty, which they linked to verbal working memory differences on 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 exposure-related factors that shape linguistic expectations and modulate the processing difficulty that readers encounter. In support of the latter approach, MacDonald and Christiansen (2002) conducted SRN simulations whereby they qualitatively fit the SR/OR RT patterns attributed to low- and high-span individuals as a function of the amount of relative clause exposure received by their networks. In addition, a human training study by Wells, Christiansen, Race, Acheson, and MacDonald (2009) documented that increased experience in reading relative clauses (compared to a control condition) altered participants' RT profiles towards matching those of ostensibly high-span individuals (and the aforementioned high-trained SRNs).

These studies imply a crucial role for statistical learning as a mediator of experience-driven effects on shaping readers' (probabilistic) expectations, thus facilitating subsequent RTs for ORs. If implicit prediction-based processes, as tapped by statistical learning mechanisms, are indeed important to such processing phenomena and sensitively reflected in prediction-task scores from our AGL-SRT paradigm, then individual differences in statistical predictive skills from Experiment 1 should systematically vary with differences in relative clause processing. Experiment 3 thus empirically tests this hypothesis using a within-subjects design.

4.1. Method

4.1.1. Participants

The last 20 participants in Experiment 1 were recruited to participate afterwards in this experiment for additional credit. Data from four of these participants were excluded (one for refusal to participate, and three due to equipment malfunction).

4.1.2. Materials

SR/OR sentence pairs from Wells et al. (2009) were used to prepare two counterbalanced, experimental sentence lists. Each list contained 12 initial practice items, 40 experimental items (20 SRs, 20 ORs), and 48 filler items. Semantic plausibility information for subject/object nouns was controlled in the experimental sentences, with comprehension questions (Yes/No format) following each sentence item.

4.1.3. Procedure

Participants were randomly assigned to an experimental list, which was presented using a standard self-paced reading, moving-window paradigm (Just, Carpenter, & Woolley, 1982). Sentence items were thus presented in random order, with both millisecond RTs for each word and accuracy for each comprehension probe recorded.

4.2. Results

Raw RTs corresponding to practice items and those in excess of 2,500 ms (1.01% of data) were excluded from analyses. RTs were length-adjusted by computing a regression equation for each participant based on the character-length of a word, and subtracting observed RT values from predicted values (Ferreira & Clifton, 1986). Means from these residual RTs were then calculated across subject- and object-relative clauses for the same sentence regions that have been analyzed in prior related work (see, e.g., Wells et al., 2009). Consistent with past studies, greater processing difficulty for ORs was reflected by substantially increased RTs at the main verb. Also in-line with prior findings, overall comprehension rate was high (86.8%, SD = 8.1), with significantly poorer accuracy observed for ORs (74.7%, SD = 19.0) than for SRs (85.6%, SD = 9.6), $t(15) = 2.66$, $p = 0.02$.

To test our hypothesis about the involvement of statistical predictive skills in relative clause processing, we correlated individuals' prediction task scores from Experiment 1 with their length-adjusted RTs at the main verb of the relative clauses, with results illustrated in Fig. 6. For SRs, there was no significant association ($r = -0.10$, $p = 0.72$), as expected, because experience has not been shown to be a factor for further facilitating processing of this comparatively easier clause-type. For ORs, however, higher prediction task scores were associated with lesser reading difficulty ($r = -0.59$, $p = 0.02$). Moreover, individual differences in prediction task scores were not predictive of RTs for any other standard SR/OR sentence regions *except*, crucially, at the main verb of ORs—the anticipated locus of observed processing difficulty. This pattern is additionally evidenced and clearly reflected in the RTs of participants when subdivided into ''high'' and ''low'' groups based on prediction task scores (with chance-level performance of 50% as the cutoff-level). As seen in Fig. 7, ''low pred'' participants ($n = 9$, $M = 42.6\%$, SD = 8.8) differed from ''high pred'' participants ($n = 7$, $M = 73.8\%$, SD = 4.8) only for processing at the critical OR main verb. The overall pattern of RTs for both SR and OR sentences closely mirrors qualitatively the pattern of ''high'' versus ''low'' experience of participants in Wells et al. (2009).

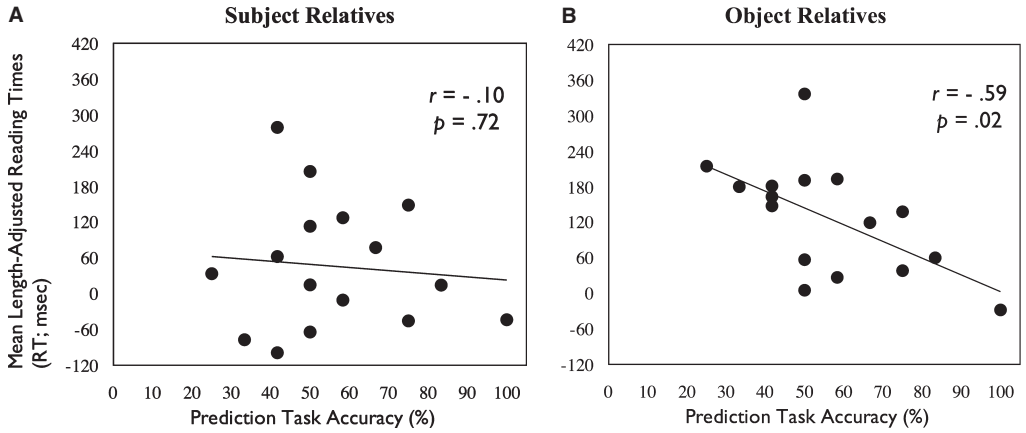

Fig. 6. Length-adjusted reading times at the main verb of subject- (A) and object-relatives (B), plotted against prediction task scores.

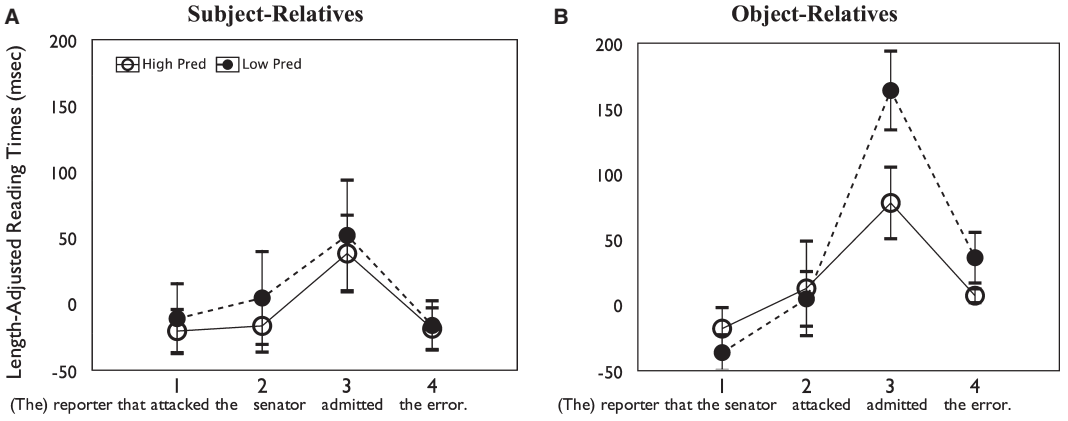

Fig. 7. Length-adjusted reading times across sentence regions of subject-relatives (A) and object-relatives (B) in Experiment 3 for participants with ''low'' (filled circles) and ''high'' (open circles) prediction task scores from Experiment 1.

These findings support the hypothesis that prediction-based processes tapped by statistical learning mechanisms (as assessed through prediction-task performances in the AGL-SRT paradigm) are substantially involved in individuals' on-line natural language processing. This conclusion is also corroborated 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 verbal working memory span scores. The current study thus expands on those findings by documenting that differences in nonadjacent statistical learning vary systematically with the on-line tracking of nonadjacent dependencies exemplified by OR sentences.

5. 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. Individual differences in performance on the statistical prediction task were shown to correlate with the processing of 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.

But how does the individual variation in statistical learning manifest itself in our AGL-SRT statistical learning task? Inspection of micro-level trajectories from Experiment 1 for good and poor statistical learners (as measured by prediction task scores) indicates distinct differences during nonadjacency learning. Thus, there are contrasts in the shape of the statistical learning trajectory, final training performance, and the response to ungrammatical items. In particular, the poor prediction-task performers do not show evidence of learning until the very end of training, contributing to the strong recovery effect on this block observable in Fig. 3. We expect that future work into such individual differences in statistical learning will benefit from closer attention to predictive processing as it unfolds over time, investigated using on-line methods such as the AGL-SRT task used here.

In broader theoretical terms, our close modeling of human performance with SRNs in Experiment 2 argues against the assumption that verbal working memory capacity operates as a basic constraint for the human results in Experiments 1 and 3; it also establishes a connection with the results from MacDonald and Christiansen (2002) in terms of common mechanisms. Their simulations with SRNs predicted that increased exposure to relative clause sentences should differentially affect ORs. Wells et al. (2009) empirically confirmed those predictions and further hypothesized that statistical learning may be centrally involved—but did not otherwise speak to what the underlying mechanisms may be. The combination of results from the three experiments reported here, however, directly supports Wells et al.'s hypothesis. In particular, not only did individual differences in statistical prediction performance correlate uniquely with on-line language processing measures at the key main verb region in OR sentences, as would be expected on an experience-based account, but prediction performance for high- and low-performing individuals on SR/OR processing also closely conformed to the pattern obtained for participants measured to have high/low verbal working memory 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 statistical learning overall is a better predictor of sentence processing skills than verbal working memory (Misyak & Christiansen, 2007), these results provide converging evidence for statistical learning as a key contributing factor to individual differences in language, and as a mechanism for producing sequential expectations for upcoming linguistic material.

Notes

1. As analyzed trials required accuracy for all three 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 it may underestimate participants' total accuracy across all single responses. For example, final-element selection accuracy across trial-types was 95.9% (2.4), 93.2% (6.5), and 94.2% (6.1).
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.

References

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 583–609.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59* (Suppl. 1), 129–164.

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

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (in press). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*.

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

Elman, J. L. (1991). Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 547–582.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505.

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

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy and relative clause comprehension. *Journal of Memory and Language*, *58*, 161–187.

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.

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass*, *2/4*, 647–670.

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: Erlbaum.

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.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, *57*, 1–23.

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. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

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

Thomas, K. M., & Nelson, C. A. (2001). Serial reaction time learning in preschool- and school-age children. *Journal of Experimental Child Psychology*, *79*, 364–387.

Van Berkum, J. J. A. (2008). Understanding sentences in context: What brain waves can tell us. *Current Directions in Psychological Science*, *17*, 376–380.

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.