On-line Measures of Prediction in a Self-Paced Statistical Learning Task

Elisabeth A. Karuza[1] (ekaruza@bcs.rochester.edu), Thomas A. Farmer[2] (thomas-farmer@uiowa.edu), Alex B. Fine[3] (abfine@illinois.edu), Francis X. Smith[2] (francis-smith@uiowa.edu), T. Florian Jaeger[1] (fjaeger@bcs.rochester.edu)

[1]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627
[2]Department of Psychology, University of Iowa, Iowa City, IA 52242
[3]Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820

Abstract

As lifelong statistical learners, humans are remarkably sensitive to the unfolding of elements and events in their surroundings. In the present work, we examined the time-course of non-local dependency learning using a self-paced moving window display. We exposed participants to an artificial grammar of shape sequences and extracted processing times, or how long they viewed each shape, over the course of the experiment. On-line learning was quantified as the growing difference in viewing duration between predictable and predictive items. In other words, as participants learned, they processed predictable items increasingly faster. Our results indicate that participants who make implicit predictions as they learn, *and* have their expectations met, achieve higher learning outcomes on an off-line post-test. Potential links between these findings, obtained with novel stimuli in an experimental context, and the role of prediction in natural language comprehension are considered.

Keywords: Statistical Learning; Adaptation; Prediction; Domain-General Processes; On-line Measures

Introduction

In order to interact effectively with our environment, it is necessary to acquire and adapt internal representations of its structure. This process is driven at least in part by the implicit extraction of statistical patterns. A substantial literature on "statistical learning" has shown empirically that learners tap into task-relevant regularities in order to segment words from continuous speech or uncover spatio-temporal relationships in visual arrays (Saffran, Aslin, & Newport, 1996; Kirkham, Slemmer, Richardson, & Johnson, 2007, respectively).

Studies of statistical learning commonly involve manipulation of a specific type of regularity: the conditional probabilities between adjacent elements. For example, in a word segmentation task, a high conditional probability between neighboring syllables might suggest that those syllables form a coherent chunk (i.e., a word). In the natural world, there are, of course, a number of inter-related regularities that learners exploit in the process of extracting structure. For one, relationships exist not only between adjacent items, but also between items that are not in direct proximity. In the case of Semitic languages, for example, many words are formed from triliteral roots in which vowels vary within fixed consonant frames. Implementing this pattern in an artificial context, Newport and Aslin (2004) demonstrated that participants could segment words from speech on the basis of probabilities between *non-adjacent* consonants (or vowels), even when the probabilities associated with *adjacent* syllables were uninformative.

Additionally, sensitivity to non-local statistical patterns has been found to induce knowledge of phrase-level grammatical relationships. Gomez (2002) has argued that the contrast between variant and invariant elements in linguistic input leads to the acquisition of non-local associations. In English, the present progressive can be formed by combining the auxiliary verb "is" and a main verb marked with inflectional suffix "ing". Thus, "is" and "ing" have a high joint probability (e.g., is eating, is sleeping, is walking, etc.), whereas the intervening main verbs vary widely. In this vein, Gomez created an artificial grammar of the form A-X-B in which pseudowords in the final position (B) were perfectly predictable given pseudowords in the initial position (A), but X items were drawn from a large set of possible elements. She demonstrated that both infants and adults acquired the non-adjacent dependencies between A and B after a period of passive auditory exposure. On the basis of these results, Gomez suggested that the presence of variability affords the ability to detect long-distance dependencies, bolstering claims that statistical learning is one plausible mechanism of grammar induction.

Employing an adaptation of Gomez's A-X-B grammar displayed in the visual modality, the present study capitalizes on a well-established behavioral metric of implicit learning, motor response time, to investigate how the allocation of processing resources during remote dependency learning changes progressively over exposure to patterned input. Importantly, nearly all studies of statistical learning assess acquisition with one off-line post-exposure test (but see, e.g., Karuza et al., 2013). As a result, we are only beginning to uncover the temporal nature of the process by which the naïve subject experiences a patterned world and derives knowledge of its underlying structure[1].

[1] Serial reaction time has long been used as an on-line measure of deterministic sequence learning. In these tasks, subjects typically respond with different fingers (i.e., motor responses) to different locations based on a visual cue that denotes the appropriate response to execute (Nissen & Bullemer, 1987). These tasks are sometimes argued to index learned associations between motor responses. Relative to standard SRT tasks, however, the present paradigm makes use of a single motor response (repeatedly

Here, we adopt a self-paced moving window display borrowed from the sentence processing literature (Just, Carpenter, & Woolley, 1982). This task enables the learner to control the rate of exposure to an artificial grammar that contains a non-adjacent dependency. Thus, we are able to collect reaction time data as participants explore and learn about a structured world. Such paradigms have previously been used to examine changes in expectations in native language (Fine, Jaeger, Farmer, & Qian, 2013), and trade on the assumption that reading times are inversely correlated with how expected the element being read is. In turn, these expectations are tied to the prior knowledge a reader brings into the task (Levy, 2008). We apply this rationale to the study of remote dependency learning in a statistical learning task. Namely, we examine the time-course of long-distance learning: as participants begin to extract structure from the input presented to them, we expect to observe a facilitation effect, a growing *decrease* in processing time, on *predictable* (B) elements relative to *predictive* (A) elements (Turk-Browne, Scholl, Johnson, & Chun, 2010; for an alternative type of prediction task see Misyak, Christiansen, & Tomblin, 2009). We seek to use on-line prediction as an index of learning, and to address the following questions:

(1) Do learners form expectations about underlying structure in the context of a novel environment? We hypothesize that the successful generation of expectations will manifest as an increasing processing benefit for predictable (B) relative to predictive (A) elements in a sequence.

(2) What types of regularities are learners sensitive to and how rapidly do they extract them? We test the hypothesis that subjects will show sensitivity to multiple types of regularity. Specifically, we examine learning of both low-level statistics (e.g., the frequency with which A and B elements occur in a given position in a sequence) and higher-level statistics (e.g., non-local dependencies between A and B). We evaluate whether the timecourse of learning depends on the complexity of the regularities present in the input.

(3) Are on-line measures of learning correlated with learning as measured on an off-line post-test? We test the hypothesis that subjects who demonstrate greater prediction effects (those with the greatest processing benefits on predictable items B) will attain higher learning outcomes as measured by a post-test.

To investigate these hypotheses, we use a measure (self-paced processing time) that remains under-explored in the context of statistical learning. This allows us to investigate the *incremental cumulative* effect of exposure (see also Hunt & Aslin, 2001). If successful, similar paradigms could be extended to investigate in more depth how learners explore the space of hypotheses about the structure of particular environments. This work also provides an important connection to research on sequential processing in more natural tasks such as prediction during spoken and written language comprehension, an issue to which we return in the Discussion (see Altmann & Mirkovic, 2009, for a review).

hitting the space bar) to examine sequential prediction of visual content while holding motor plans constant.

Materials and Methods

Stimuli

Learning was examined using an adaptation of the Gomez (2002) artificial grammar presented in the visual modality. Our experiment differs from the original Gomez (2002) study in three significant ways: (1) we presented visual shapes as opposed to recordings of spoken words; (2) we exposed an additional group of subjects to an unstructured control condition; and (3) our exposure phase was self-paced, meaning subjects controlled the presentation of stimuli during learning. Respectively, these changes enable us to test the robustness of non-adjacent dependency learning in the visual domain, to rule out item frequency or time-on-task as sources of the observed learning effects, and to observe cumulative changes in processing time as participants extract structure from the input.

Table 1. Experiment design, including ordering of tasks, number of trials, and behavioral data collected

Phase	Task	Trial *N*	Measure
1.Familiarization	Glyph matching	30	N/a
2. Exposure	A. Self-paced presentation of triplets	432	Processing time/ glyph (ms)
	B. Intermittent catch trials	144	Accuracy
3. Post-test	Familiarity judgments	12	Accuracy

Because our primary measure is processing time (PT), we took additional steps to ensure that participants were attending to the stimuli during the exposure phase. Interspersed throughout the exposure phase were 144 catch trials requiring subjects to indicate whether or not they had seen a specific item in the previous triplet sequence. Catch trials were not necessary in the original Gomez study because stimuli were presented auditorily, and processing time during learning was not an intended measure. Here, they ensured that participants actually 'read' the elements, as opposed to merely clicking through the experiment. The "shapes" in this study were actually glyphs from Ge'ez script (a writing system found in Ethiopia and Eritrea). These particular stimuli were selected because we required a large set of visually distinct items that would be unfamiliar to most native English speakers in the Rochester community.

Structured Condition Participants were exposed to a series of 3-element strings of the form A-X-B. Elements in regions A and B were drawn from a set of 6 and paired such that each A-element always co-occurred with the same B element (i.e., A1-X-B1, A2-X-B2, A3-X-B3). In contrast, X elements were drawn from a pool of 24 items. Thus, the transitional probability (TP) between *non-adjacent* items within a string (i.e., B|A) was 1.0, but the transitional probability between *adjacent* items within a string (i.e., X|A or B|X) was extremely low, only 0.04. The 3 pairs of A and B elements were combined exhaustively with the full set of X elements, rendering 72 unique sequences. As in the original Gomez study, these strings were repeated 6 times and then randomized to form a list of 432 triplets. These triplets were presented as unique trials in the exposure phase. In the subsequent test phase, participants judged the familiarity of 12 strings, half of which adhered to the A-X-B form found in the exposure phase (e.g., A1-X4-B1) and half of which violated that form because they contained unmatched A and B items (e.g., A1-X4-B3).

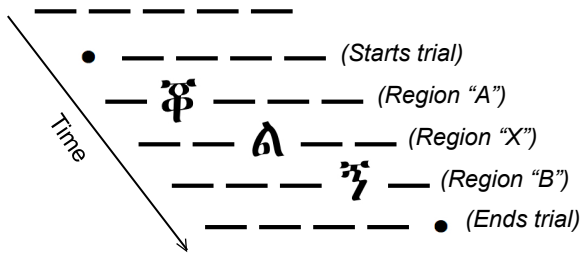

Figure 1. Example of a single triplet trial in the structured condition. Each trial began with a row of dashes. Participants advanced each item in the sequence by pressing the space bar. Response time differences between the initiation of successive elements were recorded, revealing the duration each glyph was present on the screen.

Unstructured Control To examine the effect of non-adjacent dependency learning separately from task adaptation or increasing familiarity with the visual features of glyphs, we created an unstructured control consisting of 72 non-predictive triplets repeated 6 times each. Stimuli were engineered such that the TP between any two adjacent or non-adjacent items never exceeded 0.25. Furthermore, position was uninformative in the unstructured condition; that is, A, X, and B glyphs occurred in each of the 3 presentation slots. Recall that in the structured condition, items A and B were *perfectly predictable* (TP=1.0) and always occurred in positions 1 and 3, respectively. Across conditions, individual element frequency was matched (e.g., participants always saw a total of 144 instances of glyph "A1" and 18 instances of glyph "X13"). The format of catch trials and test trials was also identical. Thus, the unstructured condition was as closely matched as possible to the structured condition, but differed along one critical dimension: the absence of a strong non-local dependency.

Participants

37 participants from the University of Rochester community are included in the present analyses. They were assigned either to the structured condition (n=19) or the unstructured (n=18) condition. All were native English speakers. They provided informed consent and were compensated at a rate of $10/hour. The experiment lasted approximately one hour, depending on the pace of the participant. Of the 42 participants who originally completed the study, 5 were excluded because their performance on the catch trials was below 70% (mean performance in remaining subjects = 90%). No participant was familiar with the glyph-based writing system used to generate the stimuli.

Procedure

The experiment consisted of 3 phases: familiarization with the individual glyphs, exposure to the structured glyph sequences, and a post-test establishing the extent of learning (Table 1). Exposure and test lists were presented in one of two orders. Procedures were identical in the structured and unstructured conditions.

Familiarization Subjects first completed a brief (~5 min) matching task. This phase ensured that any early PT effects would be reflective of learning, not of surprisal to the occurrence of a novel glyph. Each glyph was flashed on the computer screen for 2 s. Next, the participant was presented with three options and asked to select which option corresponded to the glyph they had just observed. They advanced to the next trial only after responding correctly to the current trial.

Exposure The exposure phase consisted of 432 triplets and 144 intermittent catch trials. Participants were instructed to pay attention to the screen and make their best effort on the catch trials. Regardless of condition, they were informed that stimuli might become familiar over time. There was a built-in break option every 96 trials.

The pace of the exposure trials was controlled entirely by the participant. At the start of each triplet trial, the participant saw 5 horizontal dashes centered on the computer screen. They initiated a trial by pressing the space bar, at which point the first dash became a small, opaque circle. At the next press of the space bar, the circle became a dash and the second dash was replaced by Glyph A. With another press of the space bar, Glyph A became a dash again and the next dash became Glyph X. This process continued until the trial was completed. To reduce any effects associated with initiating or ending a trial, positions 1 and 5 were always small, opaque circles. Triplet structure was embedded exclusively in positions 2-4 (Figure 1). In light of the novelty of the task, participants performed three initial practice trials consisting of number and letter, instead of glyph, sequences.

Post-test The final phase contained 12 test items. Triplets were presented in their entirety (i.e., all glyphs appeared

simultaneously). For each trial, participants indicated whether or not that sequence seemed familiar, i.e., whether they thought they had seen it in the exposure phase. In the structured condition, 6 of the test trials contained the long-distance dependency present in the input, and 6 trials contained a subtle violation of that dependency. In the unstructured condition, 6 of the test trials were seen previously in the exposure, and, similarly, 6 contained a single violation on a previously viewed triplet. Testing on the control condition allowed us to demonstrate that the extent of learning as measured at post-test is indicative of non-adjacent dependency learning, not simply explicit memory of strings presented during the exposure phase.

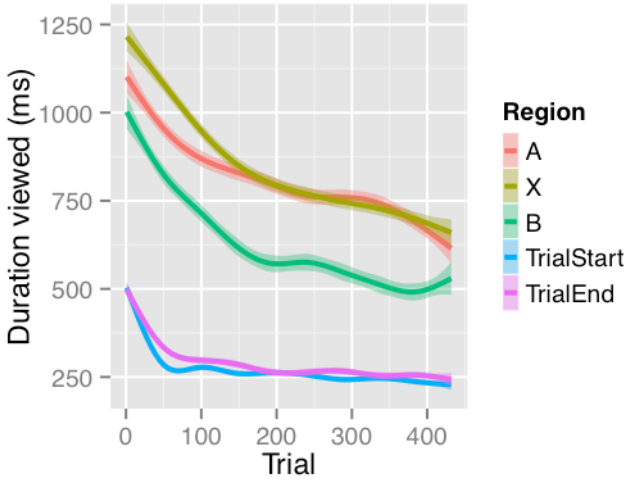

Figure 2. Smoothed estimate of mean processing time per region in the structured condition, conditional on trial. Smoothing was obtained by fitting a Generalized Additive Model to the unaggregated data, allowing for high degrees of non-linearity.

Analyses and Results

Triplets followed by an incorrect catch trial response were removed, as it was not possible in these cases to rule out subject inattentiveness on the preceding sequence (data loss 3.3%). We then excluded glyphs with a duration exceeding 6 s or falling outside 3 SDs of the mean processing time per glyph (data loss 2.5%).

Figure 2 shows 95% confidence intervals of mean processing time by trial for each Region (A, X, B, trial start, and trial end) in the structured condition. Visual inspection of the plot, supported by subsequent analyses, reveals a pronounced facilitation effect on Region B. In other words, across the course of the experiment, the third item of a sequence required less processing time relative to Regions A and X. Critically, PT on Region B began to plateau around trial 200. These data suggest that participants hit a "processing floor" midway through the experiment, at which time they were no longer afforded an additional facilitation effect by anticipating the predictable element.

The decision to include 432 exposure trials was motivated by precedent. Gomez (2002) obtained evidence of learning (as measured on post-test) after an exposure phase of this length. The present results indicate that learning, defined by the increasingly negative slope on Region B relative to Region A, began to level off by trial 200[2]. Accordingly, analyses reported below were limited to the first 200 trials. We justify our decision to subset the data in that we explicitly hypothesized an increasingly negative slope for Region B relative to Region A. While it is unlikely that learning stopped abruptly after trial 200, it is the case that slope ceases to be an effective index of learning as participants approach the processing floor. We now explore the effects of element predictability on processing time in the first 200 trials using linear mixed effects regression. To be clear, the significant interactions reported in the following sections were not obtained when these analyses were run over all 432 trials.

Protracted Learning Effects

In Model I, processing times were regressed onto all main effects and interactions of Trial (1-200), Region (B−A), and Condition (Structured−Unstructured). A second model was run over elements B and X, excluding the A elements (Model II, Region = B–X). Predictors were centered to reduce multicollinearity between main effects and interactions (fixed effect correlation rs≤0.3). Both models included random by-subject intercepts and random slopes for Region. This random effects structure was selected because (1) Trial was not a design factor and (2) adopting a more conservative random effects structure led to extremely high (rs>0.8) correlations between predictors of interest (suggesting overparameterization). Results for both models are summarized in Table 2. For each contrast, we obtained significant main effects of Trial (Model I: $\beta = -1.6$, $p<.05$; II: $\beta = -1.8$, $p<.05$) and Region (I: $\beta = -64.6$, $p<.05$; II: $\beta = -82.7$, $p<.05$), as well as a significant interaction between Trial and Condition (I: $\beta = -0.2$, $p<.05$; II: $\beta = -0.5$, $p<.05$). Unsurprisingly, subjects exhibited a general tendency to speed up over time, and they got faster in the structured relative to the unstructured condition. Notably, we found a significant three-way interaction between Trial, Region, and Condition for B relative to A (I: $\beta = -0.2$, $p<.05$). This result supports our central hypothesis, namely that with each additional trial, processing time associated with the predict*able* item should decrease more quickly than PT on the predict*ive* item, and that this difference should be greater in the structured condition. This three-way interaction was not significant for the contrast of B and X.

To evaluate the relationship between the generation of predictions about Region B and the outcome of learning, we ran an additional linear mixed effects model in which the random effect structure was specified as the by-subject slope of the interaction between Trial and Region

[2] This might suggest that post-test accuracy scores above chance would be obtained after an abbreviated learning phase.

(1+Trial*Region| Subject)[3]. We extracted these by-subject slope estimates and plotted them against post-test accuracy scores. Figure 3 reveals that change in the processing duration for Region B compared to A is significantly *negatively* correlated with post-test performance in the structured condition (r = -0.54, p<.05). That is, subjects who 'read' Region B increasingly faster than Region A tended to perform better on post-test. Participants showing the strongest prediction effects, those who generated expectations about upcoming elements and saw them met during the learning phase, performed better on the off-line measure of learning.

Table 2. Coefficients (and corresponding t-values) for each predictor. Significant values are bolded. Models I and II were run on trials 1-200. Models III and IV were run on trials 1-20.

Predictor	I (Region B−A)	II (Region B−X)	III (Region B−A)	IV (Region B−X)
Trial	**-1.6 (-25.9)**	**-1.8 (-29.7)**	**-4.8 (-2.3)**	2.2 (1.1)
Region	**-64.6 (-4.7)**	**-82.7 (-10.7)**	**-59.5 (-3.2)**	**-81.9 (-5.5)**
Condition	-49.6 (-1.4)	-33.6 (-0.8)	-37.9 (-0.9)	12.6 (0.3)
Trial*Rgn	**-0.2 (-3.5)**	0.04 (0.7)	0.1 (0.1)	**-7.3 (-3.5)**
Trial*Cond	**-0.2 (-3.1)**	**-0.5 (-7.9)**	**-8.5 (-4.0)**	-3.8 (-1.8)
Rgn*Cond	-22.2 (-1.6)	**-38.7 (-5.0)**	-2.9 (-0.2)	**-55.1 (-3.7)**
Trial*Rgn* Cond	**-0.2 (-3.3)**	0.1 (1.6)	0.5 (0.2)	**-4.3 (-2.1)**

Rapid Learning Effects

We hypothesized two aspects to learning in the context of this study: an early sensitivity to position-specific regularities and a slower extraction of non-adjacent dependencies. While the latter hinges on the learner's built-up experience with a series of subtly patterned triplets, the former should emerge after only a handful of trials. As the subject incrementally learns about the underlying process that creates the observed sequences, it follows that their initial expectations about structure should conform closely to the input. Learning of element frequency and position might then precede learning of the latent structures present in the input. Note that position 3 (Region B) in the structured condition always corresponded to one of three glyphs (B1, B2, B3). In contrast, Region X corresponded to a larger set of 24 items. To test our hypothesis that learners were sensitive to position-specific statistics early on in exposure, we evaluated the interaction of Trial, Region, and Condition in the first 20 trials. In Model III, processing times were again regressed onto all main effects and interactions of Trial (1-20), Region (B−A), and Condition (Structured−Unstructured). Likewise, a second model was run over B and X, excluding all A elements ("Model IV", Region = B−X). If participants were immediately keying into position-based statistics, then we should again observe a divergence of slopes that is more strongly negative in the structured condition. Essentially, processing time on B should speed up more quickly relative to X in the first 20 trials. We indeed found a significant three-way interaction between Trial, Region, and Condition for the contrast B−X ($\beta = -4.3$, p<.05). This same interaction was not significant for the contrast B−A. After only 20 trials, we would not expect a divergence in the slopes associated with A and B, as the learner would not have been exposed to sufficiently many instances of the long-distance dependency. That aspect of learning would require protracted exposure.

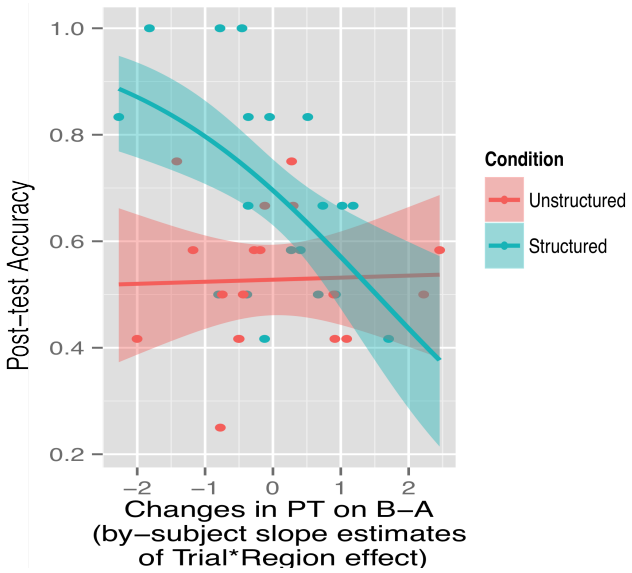

Figure 3. Relationship between change in processing time (PT) on B and performance on post-test. Chance performance for post-test is 0.5. Changes in PT for Region B compared to A are negatively correlated with post-test performance in the structured condition. Participants who sped up faster on B compared to A tended to have better learning outcomes. Participants showing the strongest prediction effects during the exposure phase achieved higher accuracy scores on the off-line familiarity judgments.

Discussion

Building on previous work examining prediction in learning (Misyak et al., 2009), we have provided fine-grained insight into the timecourse of non-adjacent

[3] This RE structure could not be used to investigate the significance of predictors, as it resulted in excessively high multicollinearity between fixed effects of interest (inflating SEs and reducing power). To investigate individual differences, however, this RE was preferred here to using the maximum-likelihood differences (means) between participants as it provided a more conservative estimate of the true between-participant differences.

dependency learning on a trial-by-trial level. Our data also contribute to small but growing literature on long-distance pattern learning that is not auditory-linguistic in nature (e.g., involving perceptually similar tones, Creel, Newport, & Aslin, 2004; or certain types of alternating visual sequences, Howard & Howard, 1997). Given that our stimuli consisted of completely unfamiliar visual tokens, Ge'ez glyphs, our paradigm is uniquely situated to probe questions of the formation of prediction when learners lack strong prior expectation about the nature of the stimuli employed.

We presented results suggesting that the processes underpinning statistical learning can be indexed by participants' ability to generate expectations about meaningful patterns and see those expectations fulfilled. Even in an artificially constructed experiment in which participants performed a fairly automatic task (repeatedly pressing the space bar), we found evidence that the brain is constantly predicting. Notably, we obtained data supporting each of our initial hypotheses: (1) Processing times revealed a progressive facilitation effect for predictable items (Region B) in the first 200 trials of exposure. This suggests that predictions, when they align with input, speed up processing of subsequent elements; (2) Analyses performed on early and protracted timecourses demonstrated that learners are sensitive to multiple sources of statistical information; and (3) Participants who made implicit predictions as they learned, and increasingly experienced their expectations being met, performed better on a post-test requiring explicit familiarity judgment. We have thus provided a link between implicit on-line and more explicit off-line measures of learning. To be clear, the correlation between these two measures does not allow us to make specific claims about the directionality of the relationship between prediction and learning. Instead, our findings serve primarily to indicate a tight coupling between the generation of implicit expectation, in this case the speed up on Region B relative to A, and a commonly used metric of learning outcome, familiarity judgments following exposure.

Recently, self-paced reading has been used to examine how expectations based on prior linguistic experience can be adapted to novel, unexpected distributions over linguistic events. Comprehenders can use prior linguistic experience to make predictions about how language is likely to be used, and those predictions are synthesized with linguistic observations in a specific environment. Fine et al. (2013) found that a priori infrequent syntactic structures, which typically incur a processing cost, are read increasingly faster in a context in which they are more probable (i.e., the structures become expected). Ongoing work considers the relationship between the learning of non-linguistic visual dependencies and syntactic adaption effects as observed by Fine et al. (2013), with the overarching goal of demonstrating that "statistical learning", as examined in artificial worlds with novel stimuli, and adaptation or priming effects in native language comprehension rely on one common, domain-general learning mechanism.

Acknowledgments

This research was supported by NSF Career Grant IIS-1150028 to TFJ, and NSF GRFs to EAK and ABF.

References

Altmann, G.T.M., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33,* 1–27.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of non-adjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1119–1130.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*, e77661.

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

Howard, J.H., & Howard, D.V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and Aging, 12,* 634–656.

Hunt, R.H., & Aslin, R.N. (2001). Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. *Journal of Experimental Psychology: General, 130,* 658–680.

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

Karuza, E.A., Newport, E.L., Aslin, R.N., Starling, S.J., Tivarus, M.E., & Bavelier, D. (2013). Neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language, 127*, 46–54.

Kirkham, N.Z., Slemmer, J.A., Richardson, D.C., & Johnson, S.P. (2007). Location, location, location: Development of spatiotemporal sequence learning in infancy. *Child Development, 78,* 1559–1571.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106,* 1126–1177.

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

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent 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.

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

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience, 30,* 11177–11187.