Tuning in to non-adjacent dependencies: How experience with learnable patterns supports learning novel regularities

Martin Zettersten (zettersten@wisc.edu)
Department of Psychology, University of Wisconsin-Madison,
1202 W. Johnson Street, Madison, WI 53706 USA

Christine Potter (cepotter@princeton.edu)
Department of Psychology, Princeton University,
220 Peretsman Scully Hall, Princeton, NJ 08540 USA

Jenny Saffran (jenny.saffran@wisc.edu)
Department of Psychology, University of Wisconsin-Madison,
1202 W. Johnson Street, Madison, WI 53706 USA

Abstract

Non-adjacent dependencies are ubiquitous in language, but difficult to learn. Previous research has shown that the presence of high variability between dependent items facilitates learning. Yet what allows learning of non-adjacent dependencies even without high variability in intervening elements? One possibility is that learning non-adjacent dependencies highlights similar structures, allowing people to learn new non-adjacent dependencies that are otherwise difficult. In two studies, we show how being exposed to learnable non-adjacent dependencies can change learners' sensitivity to novel non-adjacent regularities that are more difficult to detect. These findings demonstrate a new way in which learning can build on and shape later learning about complex linguistic structure.

Keywords: non-adjacent dependency, language learning, grammar, artificial language learning

Introduction

Non-adjacent dependencies are ubiquitous in language. For instance, English marks number agreement (e.g. *The linguists at the conference are restless*) and aspect (e.g. *Babies are learning all of the time*) via inflectional morphemes that establish dependencies between distal items. Despite their prevalence, non-adjacent dependencies in artificial grammar learning experiments are difficult to learn, both for adults and infants (e.g., Newport & Aslin, 2004; Gómez, 2002). Given their centrality to language structure, how do we learn non-adjacent dependencies that are not easily detected in the speech stream?

Previous research suggests that the input can be structured to support learners' discovery of non-adjacent regularities. For example, participants can display successful learning following extensive exposure, or when the non-adjacent dependencies are paired with correlated cues, such as related phonotactic features (e.g., Onnis, Monaghan, Richmond, & Chater, 2005; van den Bos, Christiansen, & Misyak, 2012; Vuong, Meyer, & Christiansen, 2016). In addition, the presence of high variability between dependent items facilitates learning for both infants and adults (Gómez, 2002). With inconsistent intermediate elements, learners are better able to detect the reliable associations between non-sequential elements, suggesting that surrounding information can help direct learners' attention to particular regularities.

But what allows learning of non-adjacent dependencies under less supportive circumstances? One possibility is that learners might be able to take advantage of prior experience with related structures. Previous experience can shape learners' expectations and change the statistical relations that they can track (e.g., LaCross, 2015; Lew-Williams & Saffran, 2012; Potter, Wang, & Saffran, 2017). For example, experiencing some word categories as adjacent dependencies can subsequently help learners recognize non-adjacent relationships between the same words (Lany, Gómez, & Gerken, 2007; Lany & Gómez, 2008). Yet this leaves unanswered the question of how learners might detect patterns that they only ever experience as non-adjacent regularities. In the current studies, we tested the possibility that learning one set of non-adjacent dependencies in the presence of high variability (a circumstance favorable to learning non-adjacent dependencies) later allows learners to discover novel non-adjacent dependencies they would otherwise struggle to detect.

Experiment 1

We tested whether being pre-exposed to non-adjacent dependencies in a learnable context (surrounded by high variability) would aid participants in recognizing novel non-adjacent regularities that are difficult to learn. Learners were trained on a set of artificial sentences that contained learnable, consistent non-adjacent dependencies with high variability in the intervening elements (Learnable Pre-Exposure Condition). A comparison group was trained on a set of sentences with high variability in the intervening elements, but no consistently predictable non-adjacent dependencies (Non-Learnable Pre-Exposure Condition). After this pre-exposure, all learners were trained on a new language with consistent non-adjacent relationships between a novel set of items with low variability. We predicted that

participants with pre-exposure to learnable non-adjacent relations would more readily detect novel adjacent dependencies in the exposure language.

Method

Participants

67 University of Wisconsin-Madison psychology undergraduate students (37 female; mean age: 18.8 years, *SD* = 1.04; 63 native speakers of English) participated for course credit. Participants were randomly assigned to either the Learnable (n = 32) or the Non-Learnable (n = 35) Pre-Exposure Condition.

Stimuli

All stimuli consisted of three-word sentences (e.g., aXb) with two monosyllabic words as the first and last elements (e.g., a and b) and a disyllabic word as the middle element (i.e., an X element). The items in the Pre-Exposure phase consisted of 6 novel monosyllablic words (elements a-f: *dak, tood, feep, nov, lun, kip*) and 24 novel disyllabic words (X elements: *balip, bevit, coomo, deecha, fengle, gasser, geeble, ghope, keeno, koba, lamu, loga, manu, mooper, neller, riffle, rilep, roosa, skiger, suleb, tasu, toma, vulan, wasil*). The items in the Exposure Phase were 6 new monosyllabic words (elements g-l: *pel*, *rud*, *vot*, *jic*, *bap*, *ghob*) and 3 new disyllabic words (Y elements: *kicey*, *puser*, *wadim*). In the Test Phase, there were three new Y elements to test generalization (*benez, chila, nilbo*). The items were recorded by a female monolingual speaker of English. Monosyllables and disyllables were each normalized in duration and in average intensity. The individual items were subsequently concatenated to form three-item sentences (in the form aXb, gYh, etc.) according to the Pre-Exposure and Exposure Phase Design (see Figures 1 and 2), with 100 ms of silence between each element within a sentence.

Design & Procedure

The experiment consisted of a two-part Training Phase (Pre-Exposure and Exposure) in which participants listened to a recording of the novel language. Participants' knowledge was subsequently probed in the Test Phase.

Training Phase. Participants were instructed to listen to a novel language through headphones. The training consisted of Pre-Exposure Phase and an Exposure Phase. The Pre-Exposure Phase transitioned seamlessly into the Exposure Phase, such that there was no cue to the transition except for the change in the language elements themselves. Participants viewed a series of (unrelated) natural landscape images while listening to the language.

During the *Pre-Exposure Phase* (see Figure 1), participants listened to sentences (e.g., aXb) with either consistent, learnable non-adjacent dependencies (Learnable Pre-Exposure Condition) or inconsistent non-adjacent dependencies (Non-Learnable Pre-Exposure Condition). In the Learnable Pre-Exposure Condition, participants heard elements that had one of three non-adjacent dependencies but varying middle elements (aXb, cXd, eXf). Each of the sequences in the Non-Learnable Pre-Exposure Condition used the same elements as the Learnable Condition, but recombined them such that there were no predictable non-adjacent regularities (e.g., the a element could be followed by b, d, or f with equal probability). The sentences were presented one at a time with a 750 ms silence between sentences and in one of two pseudorandomized orders with the constraint that each non-adjacent dependency (e.g., elements of the kind aXb) could occur no more than 3 times in a row. Across the pre-exposure, participants heard each sentence twice for a total pre-exposure time of 7m25s.

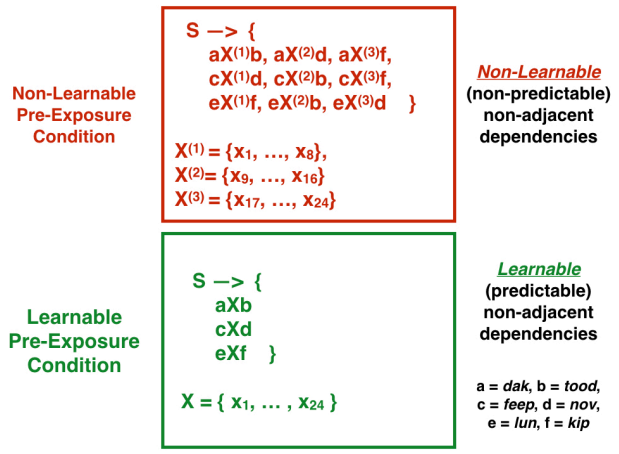

Figure 1. Pre-Exposure Phase Design.

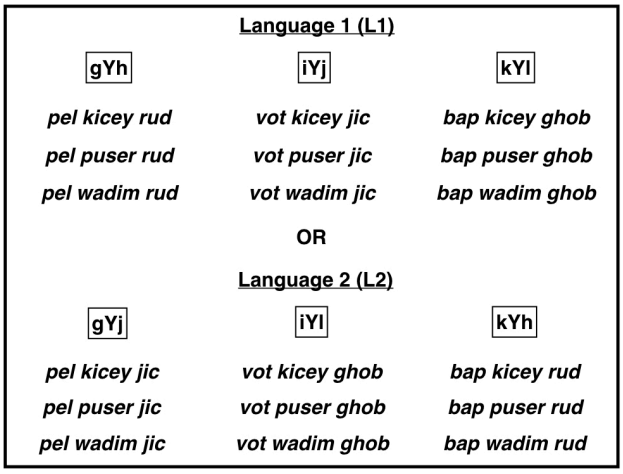

Figure 2. Exposure Phase Design

During the *Exposure Phase* (see Figure 2), participants listened to one of two possible languages (L1 or L2) with novel non-adjacent dependencies. In contrast to the Pre-Exposure Phase, there were only 3 possible middle elements. In past studies, non-adjacent dependencies of this kind with only a limited number of middle elements have proven difficult to learn (Gómez, 2002). Participants heard each

sentence 24 times across training, for a total exposure time of 11m7s. The items were presented in one of two pseudo-randomized orders for each language, with the constraint that no sentence could be presented twice in a row, and items with the same non-adjacent dependency (first and third) items could occur no more than three times in a row.

Test Phase. There were two test blocks: a *Recognition Test* block and a *Generalization Test* block (see Figure 3). In the *Recognition Test* block, participants were presented with 18 sentences one at a time in random order. For each sentence, participants were asked to decide whether the sentence matched the word order rules of the language they had just heard. Participants were also instructed that half of the sentences would match the word order rules of the language and half would not. Half of the sentences in the Recognition Test block matched sentences presented during the Exposure Phase (L1 or L2), while the other half had identical individual elements but matched the non-adjacent dependencies presented in the opposite exposure language. This counterbalancing ensured that items that were familiar in L1 were unfamiliar in L2 and vice versa.

In the *Generalization Test* block, participants judged 18 additional sentences (9 consistent with L1 and 9 consistent with L2) constructed with three new middle Y elements that were not heard during the Exposure Phase. These types of test trials were not included in the original Gómez (2002) study, but they were added here to test the degree to which participants had learned the non-adjacent dependencies between the first and last elements as opposed to simply memorizing specific sentences from the Exposure Phase.

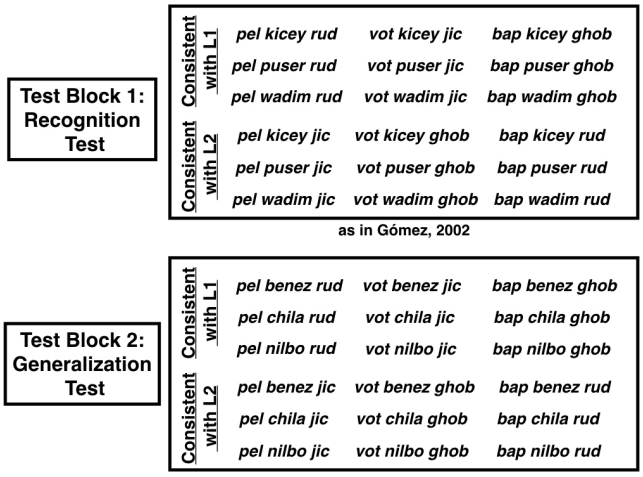

Figure 3. Test Trial Design.

Results

To test the effect of the Learnable vs. Non-Learnable pre-exposure, we predicted participants' correct responses across all test trials from Condition (centered; Non-Learnable = -0.5, Learnable= 0.5) in a logistic mixed-effects model with by-subject and by-item random intercepts. Collapsing across all test trials, participants in the Learnable Condition (M = 62.0%, 95% CI = [55.0%, 69.0%]) were more accurate overall than participants in the Non-Learnable Condition (M = 52.9%, 95% CI = [49.4%, 56.5%]), b = 0.45, Wald 95% CI = [.09, .82], $z = 2.42$, $p = .016$ (see Figure 4).

Participants in the Learnable Condition were more accurate both for Recognition Test trials (b = 0.42, Wald 95% CI = [.06, .78], $z = 2.31$, $p = .02$) and for Generalization Test trials (b = 0.41, Wald 95% CI = [.05, .78], $z = 2.20$, $p = .028$). Moreover, participants in the Learnable condition showed strong evidence of learning the non-adjacent dependency in the Exposure Phase, while participants in the Non-Learnable condition did not: in the Learnable condition, accuracy on Recognition Test trials (b = .58, Wald 95% CI = [.28, .80], $z = 4.03$, $p < .001$) and Generalization Test trials (b = .63, Wald 95% CI = [.23, 1.03], $z = 3.09$, $p = .002$) reliably differed from chance (50% accuracy), while performance in the Non-Learnable Condition did not differ from chance (Recognition Test trials: $p = .24$; Generalization Test trials: $p = .11$).

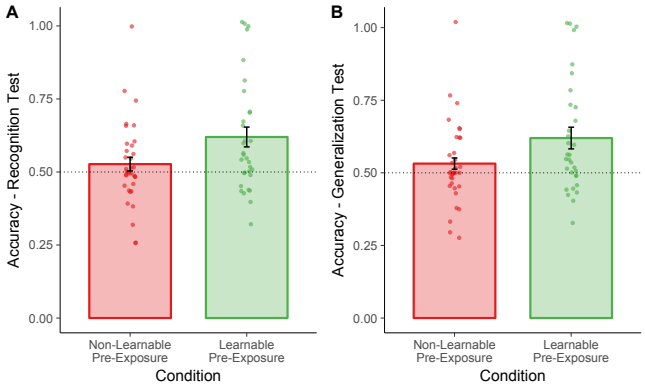

Figure 4: (A) Recognition and (B) Generalization Test Accuracy in Experiment 1. Error bars represent +1/ -1 SEs.

We also investigated the relationship between participants' performance on the two test trial types (Recognition Test vs. Generalization Test). Performance between Recognition Test trials and Generalization Test trials was correlated in both the Learnable Condition ($r = .82$, $p < .001$) and in the Non-Learnable Condition ($r = .34$, $p = .04$), though there was a significant interaction between test trial type and condition, suggesting that the relationship was stronger in the Learnable Condition, $t(63) = 3.44$, $p = .001$ (see Figure 4). Thus, participants who better recognized the sequences that they had heard during training were also more likely to demonstrate generalization of the underlying non-adjacent dependencies.

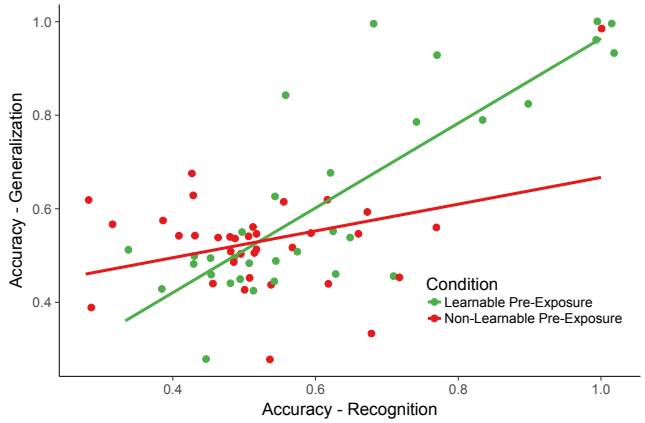

Figure 5: Experiment 1 correlation between Recognition and Generalization Test trial accuracy.

Discussion

Previous experience with consistent non-adjacent pairings with high variability in the intervening middle elements supported participants' ability to learn a new set of non-adjacent regularities that are typically difficult to learn. Participants in the Learnable Pre-Exposure and Non-Learnable Pre-Exposure conditions received *identical* experience with the target language during the Exposure phase, yet only those participants who had previously heard sequences with learnable non-adjacent dependencies demonstrated learning. These results are consistent with the hypothesis that prior experience can shape the regularities that learners are able to detect.

By testing participants' ability to generalize to new items, we found strong evidence that participants learned the non-adjacent relationship and did not simply memorize the strings that they had encountered before. Furthermore, the significant correlation between performance on the Recognition and Generalization Test trials suggests that learning was relatively robust and could be expressed in multiple ways. The stronger correlation among participants in the Learnable Pre-Exposure condition provides additional evidence that the pre-exposure experience influenced subsequent learning.

Though these results suggest that prior experience affected learning, it is not clear whether (a) the exposure to learnable non-adjacent dependencies boosted participants ability to learn new non-adjacent dependencies, (b) whether the Non-Learnable pre-exposure impeded learning of novel non-adjacent dependencies or (c) some combination of both. To address this question, we conducted a second experiment with the same conditions as Experiment 1, but with the addition of a Baseline Condition in which participants were not exposed to non-adjacent dependencies prior to the Exposure Phase. This additional condition allowed us to estimate the degree to which the pre-exposure manipulations in Experiment 1 aided or suppressed what participants typically learn about the non-adjacent dependencies in the Exposure Phase.

Experiment 2

In Experiment 2, we conducted a replication of Experiment 1 with an additional Baseline Condition in which participants received no pre-exposure experience. We predicted that there would be a linear effect between the three conditions, such that performance would be highest in the Learnable Condition, intermediate in the Baseline condition, and lowest in the Non-Learnable Condition, with significant differences between all three conditions. The linear hypothesis and analytic approach were pre-registered (Zettersten, Potter, & Saffran, 2017). A pilot study of the Baseline condition (n = 31) allowed us to estimate the approximate size of the linear effect of condition together with the data from Experiment 1 at $\eta_p = .034$.

Method

Participants

241 University of Wisconsin-Madison psychology undergraduate students (155 female; mean age: 18.5 years, *SD* = .86; 201 native speakers of English) participated for course credit. Participants were randomly assigned to the Learnable Pre-Exposure (n = 82), the Non-Learnable Pre-Exposure (n = 79), or the Baseline Condition (n = 80).

Stimuli

The stimuli were identical to the audio recordings used in Experiment 2.

Design & Procedure

The procedure for the Non-Learnable and the Learnable Pre-Exposure Condition was identical to Experiment 1. In the Baseline Condition, participants did not complete a pre-exposure phase of any kind, instead proceeding straight to the Exposure Phase. Note that as in Experiment 1, the Exposure Phase and subsequent test for learning of the non-adjacent dependencies was identical across all conditions.

Results

We fit a logistic mixed-effects model to test the linear hypothesis that non-adjacent dependency learning would improve across the three conditions (Non-Learnable < Baseline < Learnable). We followed the single contrast approach (Richter, 2015) to analyzing planned contrasts. A statistical approach that tests the residual variance in addition to the planned contrast of interest by including a second orthogonal contrast (Abelson & Prentice, 1997) leads to identical conclusions. We included Condition (coding the planned contrast as Non-Learnable: -0.5, Baseline: 0, Learnable: 0.5 to test for a linear increase across conditions) as a fixed effect and included by-subject and by-item random intercepts. There was a significant effect of Condition (b = 0.19, Wald 95% CI = [.02, .37], $z = 2.17$, $p = .030$), suggesting that there was a linear increase in performance across the three ordered conditions (see Figure 6). A similar linear effect of Condition was observed for Recognition Test trials (b =

0.21, Wald 95% CI = [.02, .40], $z = 2.18$, $p = .030$), but this effect was not significant when considering Generalization Test trials alone (b = 0.15, Wald 95% CI = [-.04, .35], $z = 1.56$, $p = .12$).

Next, we tested for differences between each condition pair by conducting pairwise comparisons, using the same modeling approach described above. As in Experiment 1, participants showed better learning in the Learnable Condition (M = 56.8%, 95% CI = [53.2%, 60.4%]) than in the Non-Learnable Condition (M = 52.6%, 95% CI = [50.5%, 54.7%]), b = .20, Wald 95% CI = [.01, .38], $z = 2.10$, $p = .036$. However, we found no significant differences between the Learnable Condition and the Baseline Condition and the Non-Learnable Condition and the Baseline Condition (M = 53.9%, 95% CI = [51.3%, 56.6%]), $ps > .18$. Accuracy reliably differed from chance across all three conditions, though the effect increased linearly from the Non-Learnable Condition (b = .10, Wald 95% CI = [.02, .19], $z = 2.45$, $p = .014$) to the Baseline Condition (b = .17, Wald 95% CI = [.06, .28], $z = 2.94$, $p = .003$) and to the Learnable Condition (b = .34, Wald 95% CI = [.16, .52], $z = 3.63$, $p < .001$).

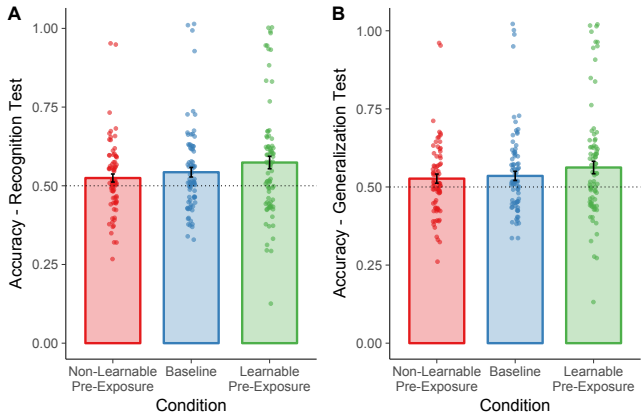

Figure 6: (A) Recognition and (B) Generalization Test Accuracy in Experiment 2. Error bars represent +1/ -1 SEs.

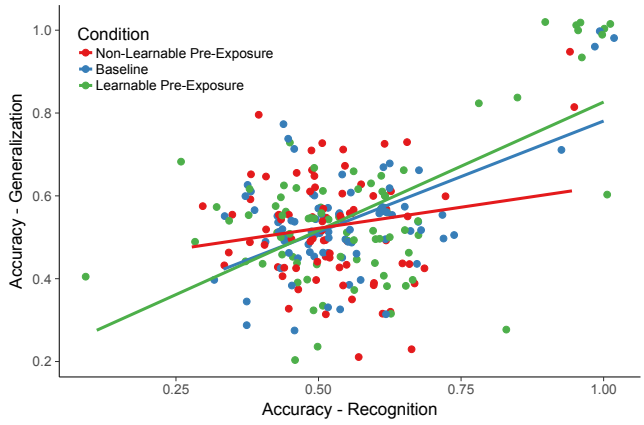

Figure 7: Experiment 2 correlation between Recognition and Generalization Test trial accuracy.

Performance between Recognition Test trials and Generalization Test trials was correlated in the Learnable Condition ($r = .62$, $p < .001$) and in the Baseline Condition ($r = .56$, $p < .001$), but not in the Non-Learnable Condition ($r = .12$, $p = .12$). There was a significant interaction between test trial type and condition (contrast coded as Non-Learnable: -0.5, Baseline: 0, Learnable: 0.5), suggesting that the correlation increased across (linearly ordered) condition, $t(237) = 2.63$, $p = .009$ (see Figure 7).

Discussion

In Experiment 2, we replicated our main results from Experiment 1 and again provided evidence that prior experience with reliable or unreliable non-adjacent dependencies can affect subsequent learning. Overall, the results followed the predicted linear pattern, with accuracy highest in the Non-Learnable Pre-Exposure condition and lowest in the Learnable Pre-Exposure condition. Thus, it appears that prior experience has the potential to both facilitate and impair later learning.

Though the main effect was consistent with our linear hypothesis, the individual comparisons between conditions were not significant. This is not entirely unsurprising, given the size of the effect for these differences. The pre-registered analysis was tuned to the size of the linear effect of condition and was perhaps too small to test for differences between the Baseline condition and the Learnable condition. Future work will test the size of the boost to later learning provided by experiencing consistent non-adjacent dependencies.

One related concern regarding Experiment 2 is that the Baseline condition has a shorter overall training phase than the two conditions that include a pre-exposure phase. Participants in the Baseline Condition may have slightly improved performance relative to participants in the Non-Learnable Pre-Exposure Condition due to less fatigue or due to experiencing less novel language material in general. Ongoing work is investigating this question by testing performance in a Baseline Condition matched to the Learnable and Non-Learnable Conditions in overall language exposure (Zettersten, Potter, & Saffran, 2018).

Another question we leave for future analyses is the existence of individual differences in participants' performance. The distribution of responses in the Learnable condition was bimodal, with a small set of participants showing perfect or near-perfect accuracy (see Figure 6). What are the characteristics of learners who are readily able to recognize non-adjacent dependencies? A question of particular interest is whether these learners show better performance on other language-related learning tasks.

General Discussion

These studies investigated a proposal for how distributional learning might build on itself, such that learners develop expectations at higher structural levels. We found that pre-exposure to learnable non-adjacent dependency structure - as compared to inconsistent non-adjacencies - differentially affected learning. Participants' learning was malleable and

susceptible to recent experience. This flexibility suggests that one way that learners can discover challenging novel regularities in language is to make use of knowledge abstracted from similar regularities.

How do the current findings this relate to infant language acquisition? Previous studies suggest that infants are sensitive to previously experienced regularities when parsing novel linguistic input (Lew-Williams & Saffran, 2012; Thiessen & Saffran, 2007). The current work supports the notion that once more difficult-to-learn structures are learned under favorable circumstances, this can support later learning. A productive next step would be to investigate the structure and variability surrounding non-adjacent dependencies that infants are exposed to early on in development. This could help uncover the extent to which infants' early experience of non-adjacent structures is shaped to bolster initial learning that infants can subsequently build on.

These results are also consistent with the view that adults' language learning is constrained by prior language experience and knowledge (e.g., Bates & MacWhinney, 1981; Seidenberg & Zevin, 2006). In these studies, participants who had experience with reliable associations were then able to detect patterns in novel materials. Likewise, adults learning a second language are better able to acquire constructions that are consistent with the regularities of their native language (LaCross, 2015). For example, native English speakers have significant difficulty with grammatical gender and often struggle to assign the correct article to a noun, but learners whose first language uses gender are more successful (Sabourin, Stowe, & de Haan, 2006). Lifelong language experience appears to encourage participants to pay attention to or ignore some structures (such as associations between articles and nouns) rather than others. Here, we provide evidence that relatively brief experience can have substantive consequences for the types of patterns to which learners are sensitive.

Acknowledgements

This research was supported by NSF-GRFP DGE-1256259 awarded to MZ and grants from the NICHD to JRS (R37HD037466) and the Waisman Center (P30HD03352, U54 HD090256). We thank Lauren Silber and Grace McCune for aiding in data collection.

References

Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypotheses. *Psychological Methods*, *2*, 315–328.

Bates, E., & MacWhinney, B. (1981). Second language acquisition from a functionalist perspective: Pragmatic, semantic, and perceptual strategies. In H. Winitz (Ed.), *Native language and foreign language acquisition (Annals of the New York Academy of Sciences, No. 379*). New York, NY: New York Academy of Sciences.

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

LaCross, A. (2015). Khalkha Mongolian speakers' vowel bias: L1 influences on the acquisition of non-adjacent vocalic dependencies. *Language, Cognition and Neuroscience, 30*, 1033-1047.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science, 19*, 1247-1252.

Lany, J., Gómez, R. L., & Gerken, L. A. (2007). The role of prior experience in language acquisition. *Cognitive Science, 31*, 481-507.

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*, 241–246.

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

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*, 225-237.

Potter, C. E., Wang, T., & Saffran, J. R. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science*, *41*, 913-927.

Richter, M. (2015). Residual tests in the analysis of planned Contrasts: Problems and solutions. *Psychological Methods*, *21*, 112–120.

Sabourin, L., Stowe, L. A., & de Haan, G. J. (2006). Transfer effects in learning a second language grammatical gender system. *Second Language Research*, 22, 1–29.

Seidenberg, M. S. & Zevin, J. D. (2006). Connectionist models in developmental cognitive neuroscience: Critical periods and the paradox of success. In Y. Munakata & M. Johnson (Eds.), *Attention and performance XXI: Processes of change in brain and cognitive development.* Oxford: Oxford University Press.

Thiessen, E., & Saffran, J. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning & Development, 3*, 75–102.

Van Den Bos, E., Christiansen, M. H., & Misyak, J. B. (2012). Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *Journal of Memory and Language*, *67*, 507–520.

Vuong, L. C., Meyer, A. S., & Christiansen, M. H. (2016). Concurrent statistical learning of adjacent and nonadjacent dependencies. *Language Learning*, *66*, 8-30.

Zettersten, M., Potter, C., & Saffran, J. (2017, November 6). Tuning in to non-adjacent dependency learning. Retrieved from osf.io/7ewmc

Zettersten, M., Potter, C., & Saffran, J. (2018, March 6). Tuning in to non-adjacent dependency learning: Experiment 3. Retrieved from osf.io/va657