Title

Characterizing the Difference between Learning about Adjacent and Non-adjacent Dependencies

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 37(0)

Authors

Wang, Felix Hao
Mintz, Toby

Publication Date

2015

Peer reviewed

Characterizing the Difference between Learning about Adjacent and Non-adjacent Dependencies

Felix Hao Wang (wang970@usc.edu), Toby Mintz (tmintz@usc.edu)
Department of Psychology, 3620 McClintock Avenue
SGM 501, Los Angeles, CA 9008 USA

Abstract

Many studies of human sequential pattern learning demonstrate that learners detect adjacent and non-adjacent dependencies in many kinds of sequences. However, it is often assumed that the computational mechanisms behind extracting these dependencies are the same. We replicate the seminal finding that adults are capable of learning dependencies between non-adjacent words (Gómez, 2002). When we eliminate the positional information about the statistical structures by embedding the structure in phrases, learners can no longer learn the dependencies. Our methods allow us to study the learning mechanisms that are more representative of the patterns in natural languages, and show that when directly compared, adjacent and non-adjacent dependencies are not equally learnable. We suggest that learning non-adjacent dependencies in language involves a different computational mechanism from learning adjacent dependencies.

Keywords: Artificial language; Non-adjacent dependencies

Introduction

Language acquisition is one of the most complex tasks that humans solve. In order to study the underlying mechanisms, researchers identify linguistic structures that serve important functions in language, and devise ways of investigating how a language learner might learn them. Distributional patterns provide structures that serve important functions, and it is crucial for language learners to learn and represent the various structures in a language.

Distributional analyses based on word sequences are viable candidates for the analyses leaners initially perform to acquire knowledge of these structures (e.g., Gómez & Gerken, 2000). For example, distributional information provides information about grammatical categories in a variety of languages (Chemla, Mintz, Bernal, & Christophe 2009; Mintz, 2003; Redington, Chater, & Finch, 1998; St. Clair, Monaghan & Christiansen 2010; Wang & Mintz, 2010; Wang, Höhle, Ketrez, Küntay, & Mintz, 2011) that appears to be used by adult and infant learners alike (Mintz, Wang, & Li, 2014; Mintz, 2006; Shi & Melançon, 2010). While other cues may provide sources for categorization, they are not as reliable cross-linguistically. For example, phonological cues vary across languages, and can even work in the opposite way across languages (for a discussion, see Mintz, Wang & Li, 2014). It is therefore crucial to understand how distributional structures are learned in terms of specifying the specific conditions under which different structures are learnable. One way to address this issue is to study how language learners acquire these distributional patterns with an artificial language learning paradigm. In these paradigms, researchers identify the important distributional structures that are present in language, and devise an artificial language with the same structures but with nonsense words. These words or syllables make up a sequence without the influence of semantics, and the design of the sequence allows inference regarding how learning happens. For example, early studies focused on adjacent dependencies, where it was shown that the transitional probability between the syllables is computed and that humans are able to represent the adjacent dependencies (e.g., Saffran, Aslin, & Newport, 1996). In these studies, infants heard syllable streams in which a given syllable either perfectly predicted the next syllable (high transitional probability) or provided no predictive power (low transitional probability). The results indicated that infants naturally chunk the elements connected by the high transitional probabilities. In other words, they represent adjacent dependencies base on high transitional probabilities, which was argued to be important for tasks such as word segmentation.

While these adjacent dependencies are important in natural language, researchers study other kinds of structures, called non-adjacent dependencies (Gómez, 2002; Newport & Aslin, 2004; Peña, Bonatti, Nespor, & Mehler, 2002). In these structures, the patterns in question concern the stable transitions between elements that are at least one element away, rather than immediately adjacent transitions. Gómez (2002) further suggested that learning non-adjacent dependencies between words occurs when the adjacent transitional probabilities are low. One way to achieve this is to have many different words in the intermediate position across occurrences of a given non-adjacent dependency.

While both adjacent and non-adjacent dependency patterns occur in natural languages, it is not clear whether the mechanisms that detect and use these types of patterns are the same, and most experiments have focused on one type of dependency or the other. However, some explicit comparisons of the learning conditions for the two types of dependencies suggest that different mechanisms are at work (Newport & Aslin, 2004; Peña et al, 2002), and that they are engaged under different circumstances. Romberg and Saffran (2013) provided some of the first evidence that learners are learning both of these types of statistical patterns concurrently. In their study, they provided learners with three-word utterances (similar to Gómez, 2002), and systematically manipulated the internal statistical structure within the three-word utterance. They showed that adult learners can readily learn both the adjacent and non-adjacent

dependencies at once, and that local dependencies influence the learning of both adjacent and non-adjacent dependencies. According to these descriptions of learning adjacent and non-adjacent dependency, the computational level difference between the two is simply the linear distance between elements that co-occur. Data from these studies suggest that participants are perfectly able to tract adjacent and non-adjacent dependencies, and the authors took this as evidence that the computational mechanism underlying the learning process is the same.

However, one feature of these studies is worth noting. In Gómez (2002) and Romberg & Saffran (2013), subjects were only exposed to three word strings in which the non-adjacent dependency involved the first and last string, which is not representative of how dependencies arise in natural languages. Specifically, both adjacent and non-adjacent dependencies are often embedded in larger sequences. We wondered whether this aspect of these artificial languages could have engaged learning mechanisms differently than if the dependencies occurred as could do in natural languages, embedded in larger sequences. In this study, we addressed this question by investigating the effects of embedding statistical patterns in other linguistic materials.

The purpose of studying the effect of embedding is twofold. First, we wanted to eliminate the positional information about the statistical structures. In the studies mentioned above, the elements involved in the non-adjacent dependencies were at sequence edges, which have been argued to engage different learning mechanisms (Endress, Nespor & Mehler, 2009). By removing the confound of the dependencies always at edge positions, we can study the learning mechanisms that is perhaps more representative of the patterns in natural languages. Second, embedding provides the extra degrees of freedom necessary to equate the statistical information in different dependency structures, so that learning can be directly compared. Furthermore, embedding makes it harder to learn structures in general. This is useful when one wishes to examine subtle differences between two learning processes, while avoiding ceiling effects.

The current experiments are set up as follows. In Experiment 1, we explore whether non-adjacent dependencies can be learned under embedded situations, providing a baseline for comparisons in subsequent experiments. In Experiment 2, we replicate a set of results from Gómez (2002) and compare them to an embedded version. In Experiment 3, we observe the effects of changing the regularities in the embedding material. Finally, in Experiment 4, we embed adjacent dependencies in sentences to allow for comparisons between adjacent and non-adjacent dependencies.

Experiment 1

In this experiment, we designed the material to be similar to the Gómez material, with one critical difference. Here we embedded the non-adjacent dependencies in other linguistic material.

Methods

Participants Twenty-four undergraduate students at University of Southern California recruited from psychology subject pool participated. Subjects were divided nearly equally into two counterbalancing conditions (see Design and Procedure).

Stimuli The stimuli were recorded by a female American English speaker (we used the same source material as in Mintz et al., 2014). The speaker pronounced one word at a time in list citation prosody, and words were digitized at 44.1 kHz for later processing. We then digitally spliced the recording into individual word files that began at the onset of each word. Word files generated from this procedure were all shorter than 0.8 seconds, so the files were padded with silence to make each file 0.8 seconds. This allowed us to concatenate word files into sentences with words occurring every 0.8 seconds. Between each artificial sentence, there was a 0.8 second pause in between to signal the start and end of each sentence.

Design and Procedure The artificial language preserved the Gómez (2002) study design with 3 frames and 6 intervening words ($A_iX_jC_i$, where i=3 and j=6). Each frame was presented 158 times, and each different intermediate word was presented 79 times with all the frames, resulting in 474 presentations in total in terms of the frame frequency. In addition, we added 1-3 words both preceding and following the AXC trigram (buffer words). The words at X position are all bisyllabic words, following Gomez (2002), whereas the buffer words are either monosyllabic or bisyllabic. Given that trigrams are surrounded by buffer words, we consider each whole phrase (front buffer words + frame + end buffer words) a sentence. The artificial material was made up such that each word occurred every 0.8 seconds, with 0.8 second of silence between all the sentences.

Buffer words consisted of 16 words that were not used in either frames position (A, C words) or the intermediate position (X words). For each sentence, a random shuffle of these words was generated, and sets of 1-3 words were selected from the list be added to the start (or end) of the sequence with the non-adjacent dependency. In other words, no words repeat within each sentence. As such, the AXC trigram could occur only after at least 1 word is presented, and the end of the trigram (C word) would not be the last word within a sentence to be heard. The buffer words under this design are randomly distributed, so there is no distributional information available in the initial and end part of the sentence that predicts the middle of the sentence.

The experiment was composed of two phases: learning and testing. The participants were asked to listen to the material to "learn a language", and they were told that they would answer questions about the language after hearing it.

In the learning phase, participants sat at a computer and listened to the stimuli through headphones. After presentation of 43 sentences without interruption, a quiz

question appeared: "What was the last word that you heard?" A numbered list of words was displayed, and participants were prompted to press a number key to respond. After the subject answered, the screen went blank and the auditory stimuli resumed. These 'quizzes' were designed to encourage subjects to attend to the material.

Testing sequences were three-word sequences, composed of either words that were consistent with the non-adjacent dependency that was in the language ($A_iX_jC_i$), or a sequence where the C word did not match the A word in the dependency ($A_iX_jC_k$, i~=k). In the test phase, a total of 12 test sequences were presented. Two languages were created, such that the correct answers in one language were the foils in the other.

Subjects heard testing sequences through headphones, and on each trial answered the question on the screen: Did you hear this sequence before? A Y/N keyboard response was collected for this question. After a short pause, the next test trial began. After the study ended, we thanked our participants and debriefed them of the purpose of the study.

Results

To assess the performance in the testing session, we coded each response as a binary variable (1 = correct, 0 = incorrect) from subjects' yes/no responses. Average proportion of correct responses was 52%. A mixed-effect logistic regression with subject as a random effect showed that performance on the test items were not different from chance ($\beta_{intercept}$=0.041, p=0.724, ns). Given that Gómez (2002) reported an effect of the number of intermediate words on the learning of non-adjacent dependencies, it is unclear whether the unsuccessful learning of the non-adjacent dependencies was due to the fact they were embedded in other words, or that there were too few intermediate words (6). In the next experiment, we address this issue.

Experiment 2

Gómez (2002) proposed that when there were only a small number of X words in the AXC structure, learners focused on adjacent rather than non-adjacent patterns, and that high variability in the intermediate position facilitates learning the non-adjacent dependencies. Although Gómez found some evidence of learning when there were only 6 X words, learning was more robust with more intervening words. Therefore, the failure to learn in Experiment 1 could have been due to the lack of variability (n=6). To test this, in Experiment 2 we replicated the Gómez (2002) study with 24 intermediate words (Experiment 2A) and then investigate the effects of embedding the sequences that have greater X-word variability (Experiment 2B).

Methods

Participants A total of 50 undergraduate students at University of Southern California in the psychology subject pool participated, half in Experiment 2A (Gómez, 2002 replication) and half in 2B (the embedded version). Two participants were excluded from the analysis because they performed below the predetermined 65% criterion in the quizzes (60%, 0%). In each version of the experiment, there were two counterbalancing conditions, such that correct test items in one condition were foils in the other (see Design and Procedure). Eleven and 13 of the participants participated in each condition. Further counterbalancing was done for the testing condition (see below for details), and subjects were further divided for that purpose.

Stimuli We used the stimuli from Experiment 1.

Design and Procedure Experiment 2A replicated the design of Gómez (2002, Experiment 1) with 24 intermediate words. As in Experiment 1, the dependencies followed an $A_iX_jC_i$ structure, with 3 A-C frames. Each frame was presented 144 times, and each different intermediate X-word was presented 6 times in each frame, resulting in 432 presentations in total in terms of the frame frequency.

In Experiment 2B, the dependency structures were the same as in 2A, but they were embedded in buffer words as in Experiment 1. Sixteen buffer words that were not any of the A, X, or C words were added to the start and end of the non-adjacent sequence, with the constraint that no words repeated within a sentence. Two languages were created, so that the correct answers in one language were the foils in the other.

The procedure was similar to Experiment 1 in that there was a learning phase and a testing phase. The learning phase followed the same procedure as Experiment 1. In contrast to Experiment 1, the testing phase (of Experiment 2A and 2B) was designed to test knowledge of both adjacent and non-adjacent dependencies. Knowledge of adjacent patterns was tested by presenting bigrams that were part of an AXC sequence (e.g., AX, or XC). Foil items were made up by presenting the reverse of the bigrams, for example, XA or CX. In order to not induce test effects, the same AX was not tested (for example, if A_1X_5 was tested, X_5A_1 was not). The choice of X words that occur in AX context and XA context was counterbalanced between subjects (the last counterbalancing step mentioned in the previous section). There were a total of 12 bigram test items. Non-adjacent dependency test items were made up similar to those in Experiment 1, where the three word sequence were either consistent from the non-adjacent dependency in the language ($A_iX_jC_i$), or not ($A_iX_jC_k$, k~=i). There were 6 non-adjacent dependency test items in total.

To avoid test effects, we tested bigrams first, then non-adjacent dependencies. If subjects were tested on the non-adjacent dependencies first, they might deduce that some of the bigrams are correct and others not by assuming that test items are informative. This deduction can be made because all the non-adjacent dependency items have the correct configuration as far as AX & XC bigrams are concerned. Because these bigram tests are constructed differently (positional changes) from the non-adjacent dependency test items (co-occurrence change) and the bigram test in

Experiment 4, direct quantitative comparisons can only be made for adjacent tests between Experiments 2A, 2B and 3.

As in Experiment 1, subjects listened to testing sequence through headphones and responded (yes or no) to the test questions via computer keyboard.

Results

Experiment 2A successfully replicated Gómez (2002): Participants learned the non-adjacent dependencies (M=71.4%, $\beta_{intercept}$=0.916, p=0.0003). Participants also successfully learned the adjacent dependencies (M=76%, $\beta_{intercept}$= 1.15, $p=9*10^{-6}$). We therefore found no evidence that attention to the non-adjacent patterns was triggered by difficulty in remembering the adjacent sequences. These findings are consistent with those in Romberg & Saffran (2013), where subjects demonstrated simultaneous learning of adjacent and non-adjacent dependencies.

Recall that the dependency structures in Experiments 2A and 2B were identical; the only difference between the experiments was in the embedding of the dependencies in 2B. However, in contrast to 2A, participants in 2B did not show evidence of learning the non-adjacent dependencies (M=52.8%, $\beta_{intercept}$=0.111, p=0.505, ns). Furthermore, we found no evidence that participants learned the adjacent dependencies either (M=52.8%, β=0.111, p=0.346, ns).

Taken together, these results suggest that embedding non-adjacent dependencies hinders successful learning of adjacent and non-adjacent dependencies. The same patterns that were successfully learned when they were presented in isolation were apparently not-learnable when surrounded by other words.

One factor in the embedding version of this experiment (Experiment 1 and Experiment 2B) is that the buffer words in which the dependency structures were embedded were uniformly random and did not follow a grammar. This means that there was no reliable statistical information in these parts of the language. Given that the first few words of most sentences are buffer words, subjects may have simply "tuned out" when there was no discernable pattern to be found, disengaging the mechanism that would typically learn the dependency patterns (Gerken, Balcomb, & Minton, 2011). Experiment 3 was designed to address this question.

Experiment 3

In Experiment 3, we modify our embedding of non-adjacent dependencies by making the buffer words appear in a fixed sequence.

Methods

Participants A total of 24 undergraduate students at University of Southern California in the psychology subject pool participated. Half of the participants participated in each condition, and further counterbalancing of the bigram testing was done by evenly dividing the subjects in the same condition, similar to Experiment 2B.

Stimuli We used the same word stimuli as in Experiment 2B.

Design and Procedure Experiment 3 differed from Experiment 2B only that instead of the buffer words occurring in random order, they now adhered to a consistent order. For example, when there was only one sentence initial buffer word, it was the same word each time. When a sentence started with two words, it was always the same two words (that did not include the buffer word that only occurred as a singleton buffer word). In this way, the transitional probability within the buffer portion of the sentences is kept at 1; the transitional probability between the buffer word and the A word, the first word in AXC structure is 1/3 because each word that immediately precedes any A word precedes all of them in equal proportion.

Two languages were created, so that the correct answers in one language were the foils in the other.

Results

As in Experiment 2B, participants showed no evidence of learning the non-adjacent dependencies in this embedding condition (M=53.5%, $\beta_{intercept}$=0.139, p=0.4, ns). Furthermore, there was no evidence that participants learned the adjacent patterns either: (M=52.8%, $\beta_{intercept}$=0.111, p=0.346, ns). Thus, providing predictable patterns in the buffer material did not make the embedded dependencies easier to learn.

In light of the findings so far, it is important to note that the frequencies of adjacent and non-adjacent patterns in these languages are very different. For example, in Experiment 3, each non-adjacent dependency occurred 144 times, whereas each adjacent dependency occurred only 6 times. It is possible that there were different reasons why adjacent and non-adjacent dependencies were not learned. In the case of adjacent dependencies, the frequency of the bigrams may have been too low for the pattern to have been detected and remembered. Experiment 4 addresses this.

Experiment 4

In Experiment 4, tested only adjacent dependencies, making them statistically equivalent to the non-adjacent dependencies in Experiments 1-3.

Methods

Participants Twenty-five undergraduate students at University of Southern California in the psychology subject pool participated. Twelve and 13 of the participants participated in each condition.

Stimuli We used the same word stimuli we used in Experiment 1.

Design and Procedure Experiment 4 was based on Experiment 2B, except the middle X position in the $A_iX_jC_i$

structure was removed, making the former non-adjacent pattern adjacent. Each A_iC_i sequence occurred 144 times.

Two languages were created, so that the correct answers in one language are the foils in the other.

In the testing session, we used 6 test items, similar to the non-adjacent tests in previous experiments. These 6 items consisted of 3 correct bigram pairs (A_iC_i), and 3 foil bigram pairs (A_iC_k, k~=i).

Results

Mixed-effect logistic regression revealed that participants learned the bigrams successfully (M=60.7%, $\beta_{intercept}$=0.433, p=0.01). The fact that there were only 3 high frequency adjacent dependencies apparently induced learning, even though the surrounding buffer words were completely random and unpredictable.

Discussions

In a series of experiments, we explored the effects on learning of embedding adjacent and non-adjacent dependency patterns within a larger sequence of words. Our study is the first we know of that contrasts embedded non-adjacent dependency with embedding adjacent dependencies with language learning (see Van den Bos & Christiansen, 2009 for data from symbols sequence learning). While both the structured embedding (Experiment 3) and unstructured embedding (Experiment 1 & 2B) yielded no detectible learning of non-adjacent dependencies, embedding bigrams within larger random sequences did not impede their detection when they were very frequent (Experiment 4), but did when they were less frequent (Experiment 2A vs. Experiment 2B). Thus, for non-adjacent dependencies, the alignment of one or both of the dependent entities with edge positions may be important (Endress et al., 2009; primacy & recency effects, Deese 1959).

The present findings raise questions for theories of language acquisition. If adult learners cannot extract non-adjacent dependencies when they are embedded within utterances, does this mean that those dependencies cannot be learned from distributional analyses? In evaluating the implications of these findings for theories of grammatical acquisition, it is important to consider other ways in which these artificial languages differ from natural ones. One way is that the utterances used here do not implement natural language prosody. We have preliminary evidence that placing a prosodic contour on the utterances may facilitate learning non-adjacent dependencies in embedded materials (Reddy, Wang & Mintz, in prep). The continuous nature of a prosodic contour may focus leaners on relations between items in the contour, especially non-adjacent ones.

A related difference between these artificial languages and natural ones is in the timing of words. In these experiments, utterances were concatenated words with brief intervening pauses, such that there was always 0.8 s between word onsets. This is a relatively slow rate of speech, with unnatural timing characteristics. It is conceivable that this mode of presentation makes detecting patterns of non-adjacent elements more difficult because they are not temporally close. Future studies are needed to determine the degree to which these stimulus properties could influence non-adjacent pattern detection.

Finally, it is worth noting that partially embedded non-adjacent dependencies in which dependent items sometimes occur at edges enables the detection of non-adjacent patterns (Mintz et al., 2014; Reeder, Newport & Aslin, 2013). It is possible that having exposure to elements at edge positions may be necessary for initially detecting non-adjacent dependencies, but that once a pattern is recognized, it can be detected in fully embedded contexts as well. We leave this question for future research.

It is also possible that the Gómez paradigm adopted here may not be well suited for testing implicit learning. Statistical learning is often characterized as tapping into implicit learning (Saffran, Newport, Aslin, Tunick, & Barrueco 1997, Turk-Browne, Jungé, & Scholl 2005). Studies on implicit learning suggest that the learning process does not require explicit instructions, and the representations resulted from learning can be probed with implicit measures. Vuong, Meyer, and Christiansen (in press) used an SRT task to measure sequence learning, and it would be important to see if the same pattern from the motor domain shows up in language learning. Asking yes/no questions about whether particular phrases are in a language requires some explicit representation for the phrases, and this may not be the most relevant way of testing whether implicit learning is successful. Indeed, many participants in our experiments answer yes to all the questions, which attribute to the null results. Yet, analyzing the data only with individuals who showed variability in their responses does not qualitatively change the pattern of the data. We are working on new measures of non-adjacent dependency learning that do not require testing explicit representations. However, what our findings clearly show is that when subjects learn non-adjacent dependencies, they also detect patterns in adjacent items, consistent with findings in Romberg & Saffran (2013). What we seem to have shown is that whatever mechanism is robustly detecting adjacent dependencies is not operating over a wider range of input. It is an open question (at least from our data) whether that means that when learners do detect non-adjacent relationships that a different mechanism is engaged, or whether it is the same mechanism that is guided by additional information (e.g., edges, etc.). This is a question for future research.

An alternative explanation for our findings is that detecting non-adjacent patterns is simply harder (e.g., due to the additional degrees of freedom compared to adjacent patterns) and that embedding the patterns made their detection even more difficult given relatively brief exposure to the language, but with more exposure, subjects would detect the patterns. We cannot rule out the possibility that more exposure would have lead to successful learning. However, in Experiment 3, the embedded contexts did not add a large amount of complexity since the buffer patterns

were highly predictable, and yet subjects did not detect the dependencies. At minimum, these findings suggest that learners' ability to detect non-adjacent patterns is highly constrained.

In conclusion, this study shows that one should be cautious about the conclusions one draws regarding learning in artificial languages when the patterns in question are made prominent. A motivation behind using artificial languages is that one can design the language to focus on particular mechanisms of interest. Although this can be an extremely useful and fruitful approach, the process of simplifying can change the learning problem in unintended ways. We are by no means the first to raise this issue, but here we have provided evidence of one way in which (perhaps) seemingly peripheral design considerations could have important consequences. But this situation has also given rise to the insight that learning adjacent and non-adjacent lexical patterns may engage different mechanisms that are sensitive to different kinds of information, as has been proposed for patterns within words (Endress et al, 2009; Peña et al., 2002; but see also Pacton & Perruchet, 2008). Taken together, these results suggest that the mechanism for learning non-adjacent lexical dependencies is more nuanced than previously believed. While adjacent dependencies can be learned in embedded context, learning non-adjacent dependencies is very sensitive to the details of the context, and may involve factors beyond mere statistical regularity.

References

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, 12(3), 396-406.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of experimental psychology*, 58(1), 17.

Endress, A.D., Nespor, M. & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, 13(8), 348-353.

Gerken, L., Balcomb, F. K., & Minton, J. L. (2011). Infants avoid "labouring in vain" by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*, *14*(5), 972–979.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431-436.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs*, 31-63.

Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. *Cognitive psychology*, 75, 1-27.

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*(1), 80–96.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. Science, 298(5593), 604-607.

Reddy, K., Wang, F. H. & Mintz, T. H. (in prep). Prosodic contours help the learning of non-adjacent dependencies.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30–54.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4), 425-469.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science, 37*(7), 1290-1320.

St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: flexible frames for language acquisition. *Cognition*, *116*(3), 341–360. doi:10.1016/j.cognition.2010.05.012

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996).Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental Language Learning: Listening (And Learning) out of the Corner of Your Ear. *Psychological Science, 8*, 101-105.

Shi, R., & Melançon, A. (2010). Syntactic Categorization in French-Learning Infants. *Infancy*, 15(5), 517-533.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*, 552-564.

Van den Bos, E., & Christiansen, M.H. (2009). Sensitivity to nonadjacent dependencies embedded in sequences of symbols. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, Amsterdam, p2068.

Vuong, L.C., Meyer, A.S. & Christiansen, M.H. (in press). Concurrent statistical learning of adjacent and nonadjacent dependencies. *Language Learning*.

Wang, H., Höhle, B., Ketrez, N. F., Küntay, A. C., & Mintz, T. H. (2011). Cross-linguistic Distributional Analyses with Frequent Frames: The Cases of German and Turkish. In N. Danis, K. Mesh, & H. Sung (Eds.), *Proceedings of the 35th annual Boston University Conference on Language Development* (pp. 628-640). Somerville, MA: Cascadilla Press.

Wang, H., & Mintz, T. H. (2010). From linear sequences to abstract structures: Distributional information in infant-direct speech. *In Proceedings Supplement of the 34th Boston University Conference on Language Development.*