All Together Now: Concurrent Learning of Multiple Structures in an Artificial Language

Alexa R. Romberg, Jenny R. Saffran

Department of Psychology and Waisman Center, University of Wisconsin

Abstract

Natural languages contain many layers of sequential structure, from the distribution of phonemes within words to the distribution of phrases within utterances. However, most research modeling language acquisition using artificial languages has focused on only one type of distributional structure at a time. In two experiments, we investigated adult learning of an artificial language that contains dependencies between both adjacent and non-adjacent words. We found that learners rapidly acquired both types of regularities and that the strength of the adjacent statistics influenced learning of both adjacent and non-adjacent dependencies. Additionally, though accuracy was similar for both types of structure, participants' knowledge of the deterministic non-adjacent dependencies was more explicit than their knowledge of the probabilistic adjacent dependencies. The results are discussed in the context of current theories of statistical learning and language acquisition.

Keywords: Statistical learning; Language acquisition; Implicit learning; Distributional information

1. Introduction

Language acquisition is one of the most complex tasks facing human learners. Natural languages contain many levels of structure, from the distribution of phonemes within words to the distribution of words within phrases and phrases within utterances. Humans master these structures in their native language early in childhood, suggesting that they are powerfully attuned to language structure. Simplified artificial languages provide useful models to test how sensitivity to distributional structure contributes to natural language learning. Studies using artificial languages have revealed that human infants and adults readily absorb structure from language-like stimuli, including speech sound categorization (Maye, Werker, & Gerken, 2002), word segmentation (e.g., Saffran, Aslin, & Newport,

1996; Saffran, Newport, & Aslin, 1996), lexical categories (e.g., Lany & Saffran, 2010; Mintz, 2002), or simple grammatical structure (e.g., Gómez, 2002; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Saffran, 2002). These studies (and many others) have formed the foundation for theories that posit that learning the distributional properties of language is central to language acquisition (for recent reviews, see Aslin & Newport, 2012; Romberg & Saffran, 2010).

While many studies have examined learners' abilities to track distributional structure in artificial languages, most have focused on the acquisition of one type of structure at a time. For example, studies investigating word segmentation based on conditional probabilities have used materials with dependencies between adjacent syllables (Aslin, Saffran, & Newport, 1998) or non-adjacent segments (Newport & Aslin, 2004) or syllables (Endress & Bonatti, 2007; Peña, Bonatti, Nespor, & Mehler, 2002); studies investigating phrase structure learning based on conditional probabilities have again used dependencies between adjacent words (Saffran, 2002; Thompson & Newport, 2007) or non-adjacent words (Gómez, 2002; Gómez & Maye, 2005), but not both. Other studies have asked whether initial exposure to adjacent relationships can bootstrap subsequent acquisition of non-adjacent relationships when intervening items are added (Lany & Gómez, 2008). These types of designs are important, because they test learners' sensitivity to particular types of distributional information. However, natural languages contain multiple levels of structure with dependencies between both adjacent and non-adjacent elements of an utterance, such as in relative clauses or "frequent frames."

To successfully comprehend utterances with a relative clause, such as "*The dog that chases the cat is barking*," listeners must track a non-adjacent dependency (both semantic and grammatical) across an intervening phrase that is itself semantically and grammatically linked to the initial phrase. In this example, *is barking* refers to *the dog* and the grammatical number of *dog* and *is barking* must agree, constituting the non-adjacent dependency. Additionally, the relative clause *that chases the cat* gives more information about *the dog* and also must agree in number, constituting the adjacent dependency. Interestingly, recent work has revealed relationships between adults' ability to learn non-adjacent dependencies in an artificial language (using a serial response task) and their speed of processing relative clauses in English (Misyak, Christiansen, & Tomblin, 2010), as well as relationships between learning of adjacent dependencies (in visual and auditory grammar learning) and speech processing (Conway, Bauernschmidt, Huang, & Pisoni, 2010). These connections suggest that laboratory tasks investigating learning of statistical structures are tapping skills relevant for natural language use.

Another aspect of natural language that involves non-adjacent structures is the "frequent frame": a phrase that differentially co-occurs with words in different lexical categories (e.g., nouns and verbs; Mintz, 2003, 2006). For example, in the frame "*was X-ing*," X is likely to be a verb (as in "*The girl was playing*"), whereas in the frame "*the X is going*," X is likely to be a noun. These frequent frames contain dependencies between both adjacent and non-adjacent elements. In the phrase *was X-ing*, *was* predicts *–ing*, forming the non-adjacent regularity of the frame. Additionally, because X is likely to be a verb, there is an adjacent regularity between *was* and X (i.e., the conditional probability

of *run* given *was* is higher than that of *cat* given *was*). The usefulness of frequent frames as a cue to lexical categories depends on the distributional properties of the language, and frequent frames are not equally useful for all natural languages (Stumper, Bannard, Lieven, & Tomasello, 2011; see also St. Clair, Monaghan, & Christiansen, 2010 for discussion).

Mintz (2002) demonstrated that adult learners are sensitive to the distributional patterns of frequent frames. Participants listened to an artificial language in which a group of four X words was heard in each of three A_B frames (the X word fell between the A and B words). Three of these X words were also heard in a fourth A_B frame. Participants were then asked whether particular three-word strings had occurred in the language they listened to, and familiarity scores were derived from accuracy and confidence ratings. Familiarity scores were the same for AXB strings actually heard during familiarization and for the AXB string omitted from familiarization, suggesting that participants had categorized the four X words together. This interpretation implies that participants noticed the A_B frames (or the AX or XB conditional probabilities) and used them to categorize the words. However, the study was designed to test only the categorization of X words; there was no direct test of whether participants had learned the A_B non-adjacent regularities or the AX or XB adjacent regularities.

How learners track these multi-layered structures remains unknown. Prior studies have been designed to detect learning of only a single type of distributional structure. To avoid confounds, other dependencies are typically held constant, making them uninformative about the structure of the language. For example, the language employed by Gómez (2002) to test non-adjacent dependency learning consisted of three-word phrases in an AXB structure. The X words were evenly distributed between the A_B frames in all conditions, so that the adjacent dependencies were flat by design—all X words were equally likely to follow each A word and to precede each B word. Gómez found that adults and 18-month-old infants were more successful at learning the non-adjacent structure when there were many X words in the language, discriminating between familiar and novel strings when there were 24 X words, but not when there were 12. As the number of X words increases, the conditional probabilities between adjacent words drop, while the non-adjacent probabilities remain high. Gómez hypothesized that the low conditional probabilities between adjacent words served to shift participants' attention to the highly regular non-adjacent frames. This pattern of results raises important questions about how learners represent the adjacent and non-adjacent regularities. Does a highly regular sequential structure dominate learning so that participants no longer track more subtle dependencies? Must learners attend to only one layer of structure at a time, or can they switch between them flexibly? When the input has multiple layers of structure, do all learners begin by attending to the same level of analysis or are there different learning trajectories that might lead to mastery?

This study was designed to investigate these questions by employing an artificial language in which we manipulated the strength of both adjacent and non-adjacent dependencies. We used an AXB language based on Gómez (2002) with high variability in the X position. However, rather than distributing the middle X elements evenly across all

frames, subsets of X words were more frequent in some non-adjacent frames than others. We tested adult participants to determine whether they learned the non-adjacent and/or adjacent structure of the language. We also asked them to rate their confidence in their answers. Confidence ratings provide additional information about participants' knowledge and allow us to ask whether participants are equally aware of the different types of patterns built into the artificial language. Confidence ratings are not always related to knowledge (e.g., Dienes, Altmann, Kwan, & Goode, 1995) and testing for correlations between confidence and accuracy allows us to investigate whether participants have similar explicit access to their knowledge of different levels of linguistic structure.

To investigate the trajectory of learning (i.e., whether participants tended to learn one structure before the other), we tested participants after different lengths of familiarization with the artificial language. In Experiment 1, we used a language in which the adjacent regularities were strongly highlighted by their statistical properties, and in Experiment 2, we used a language in which the adjacent regularities were less prominent. Across both experiments, the artificial languages contained deterministic non-adjacent regularities and probabilistic adjacent regularities. If participants initially track adjacent structure and their attention is only shifted to non-adjacent structure after they detect high variability (i.e., low conditional probabilities) among adjacent relationships, we would expect that learning of the non-adjacent structure would only emerge with longer exposures to the language. However, if the deterministic non-adjacent frames are highly salient, they may be learned early in exposure, with sensitivity to the probabilistic adjacent structure only developing with time. Finally, it is possible that individual learners do not track both types of structure, but instead learn only adjacent or non-adjacent regularities, with the relative salience or strength of the statistical cues determining what is learned.

2. Experiment 1

2.1. Method

2.1.1. Participants

A total of 156 participants (26 per condition) were recruited from the undergraduate research pool at the University of Wisconsin—Madison; all reported normal hearing and received extra credit for their participation. An additional 18 participants were tested but excluded from the analyses due to missing three or more of the Screening test items (5; see below); failure to respond to multiple test items (5); and experimenter or technical error (8).

2.1.2 Materials

Following Gómez (2002), participants heard a list of three-word phrases of the form AXB. The A words were *pel, vot,* and *dak* and the B words were *jic, rud,* and *tood.* There were 16 X words: *balip, benez, deecha, fengle, gensim, gople, hiftam, kicey, loga, malsig, plizet, puser, roosa, skiger, suleb,* and *vamey.* Phrases were recorded by a female

native English speaker at a slow pace with lively intonation. One token of each word was used throughout the familiarization and test. Tokens were spliced together with 250 ms silence between words and 750 ms silence between phrases. The duration of each three-word phrase was approximately 3 s.

Four counterbalanced languages were constructed (the complete languages and test items can be found in the Appendix). The basic form of each language was as follows: Each A word was paired with a B word to form a frame in which the A word perfectly predicted the B word (the non-adjacent structure). To create the adjacent structure, the 16 X words were separated into four groups of four words. The *Evenly Distributed* (X_{ED}) words were the same for each frame and occurred with equal probability in each frame. The remaining three groups were assigned a probability relative to each A_B frame: *Higher Probability* (X_{HP}), *Lower Probability* (X_{LP}), or *Unattested*. For each X word, there was one frame in which it was X_{HP}, one frame in which it was X_{LP} and one frame in which it was unattested. For example, *loga* was X_{HP} in the *dak_tood* frame, X_{LP} in the *vot_jic* frame, and unattested in the *pel_rud* frame. Within each frame, X_{HP} words were four times more frequent than X_{LP} words and X_{LP} and X_{ED} were equally frequent (see Fig. 1 for the bigram probabilities).

This design rendered a language of trigrams in which the relationship between the first and third words was deterministic and the relationship between the first and second words (and second and third words) was both probabilistic and constrained by the frame. We created a corpus of 72 strings that reflected the probability structure depicted in Fig. 1. The corpus consisted of 4 repetitions of each $AX_{HP}B$ string and 1 repetition of each $AX_{LP}B$ and each $AX_{ED}B$ string, for a total of 24 repetitions of each A_B frame. The corpus was randomized with the constraint that no AXB string was repeated immediately.

The uneven distribution of X words embedded within frames meant that some X words were informative as to the frame they are in. Using the example introduced above, if the current word is *loga*, the previous word was more likely *dak* than *vot,* and was definitely not *pel*. Likewise, the next word is more likely *tood* than *jic* and is definitely not *rud*. However, not all X words were informative—the four words in the X_{ED} group occurred with equal probability in each frame, so that *gople* was equally likely to follow *dak, vot*
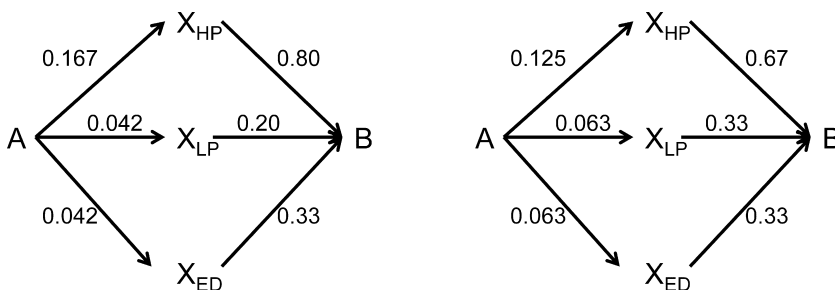

Fig. 1. The structure of the language used in Experiment 1 (left) and Experiment 2 (right). The numbers represent the bigram transitional probabilities for specific A, X, and B words.

or *pel* and to be followed by *tood, jic* or *rud*. Though this language is more complex than one with only non-adjacent dependencies, the fact that some of the X words were linked to specific frames may actually help learners detect the non-adjacent dependencies by providing an informative context (Cleeremans, Servan-Schreiber, & McClelland, 1989).

We exploited the differences between the groups of X words to create a test that probed the acquisition of both the non-adjacent and the adjacent structures. The Non-adjacent test items determined whether participants could distinguish A_B frames belonging to the familiarization language from A_B frames that never occurred during familiarization (e.g., *pel_rud* vs. *pel_jic*). The X_{ED} words were used in these test items because they were heard equally often in each frame during familiarization and therefore the adjacent probabilities were uninformative for choosing between the test strings (e.g., for the test item *pel gople rud*, the bigrams *pel gople* and *gople rud* were heard with equal frequency regardless of whether *pel_rud* was a legal frame in the familiarization language). Thus, the participants' judgments could be based only on the non-adjacent structure.

The Adjacent items tested whether participants had learned which X words were likely to appear in each frame. For these items, participants had to choose between a string with a legal frame containing an X_{HP} word for that frame and a string with a different legal frame containing an X word that was unattested in that frame. Importantly, each of the X words used in these test items occurred equally often during familiarization. Thus, correct judgments must be based on recognizing the frame in which the X word appeared during familiarization—either by recognizing the trigram as a whole or by attending to the adjacent conditional probabilities within the trigram.

Finally, we included Screening items designed to identify participants who did not pay attention during familiarization. These items contrasted familiar AXB strings with AXB strings containing novel X words and illegal A_B frames. Participants who missed three or more Screening items were excluded from the analyses and Screening items were not analyzed further.

Each of the four counterbalanced languages contained the same set of A, B, and X words. The languages and test items were constructed so that the correct answer for each test item would be one phrase for half the participants and the other phrase for the remaining participants. Languages 1A and 2A had identical X distributions but different A_B frames. Languages 1A and 1B had identical A_B frames but different distributions of X words between those frames (the same was true for 2A and 2B). Participants who were familiarized with Language 1A shared their Non-adjacent test items with participants from Language 2A (i.e., each Non-adjacent test item had one string with a frame from Language 1A and one string with a frame from Language 2A) and their Adjacent test items with participants from Language 1B (i.e., each Adjacent test item had one string with an X word that was HP in 1A but unattested in that frame in 1B and one string with an X word that was HP in 1B but unattested in that frame in 1A). Likewise, participants familiarized with Language 2B shared their Non-adjacent test items with participants from Language 1B and their Adjacent test items with participants from Language 2A. The 156 participants were randomly assigned to one of the languages in approximately equal numbers (1A = 41, 1B = 39, 2A = 36, and 2B = 40). There were

at least six participants assigned to each language within each of the six conditions determined by Exposure Duration and Test Order ($N = 26$ per condition, see below).

2.1.3. Procedure

We manipulated exposure length across participants to investigate the dynamics of learning. The shortest exposure (7 min) consisted of two runs through the corpus for a total of 144 phrase tokens; the medium exposure (14 min) included four runs through the corpus for a total of 288 tokens; and the longest exposure (21 min) included six runs through the corpus for a total of 432 tokens. Note that the longest exposure is the same as that used by Gómez (2002).

Stimulus presentation and data collection were performed with E-Prime 2.0 Professional (Schneider, Eschmann, & Zuccolotto, 2002). Participants were told they would be listening to a nonsense language and that afterward they would be asked questions about the language. They were asked to attend to the language but that they were not required to memorize or actively figure out anything about the language. Participants sat at a computer workstation and listened to the familiarization audio file over Koss UR/29 headphones (Koss Corporation, Milwaukee, WI, USA). After completing familiarization, the experimenter described the test procedure. Each test question consisted of two three-word-phrases separated by 1,000 ms. Participants were asked to indicate which phrase was most similar to the familiarization language via key-press. After they made their choice, they were asked to indicate how confident they were in their answer using a scale from 1 to 7, with 7 being very confident and 1 being very unsure. Participants had 10 s to choose the correct phrase and then unlimited time to make their confidence rating.

There were 18 test questions, 6 each testing the Non-adjacent and Adjacent structures and 6 Screening items. Test questions were blocked by question type and the order of the blocks was treated as a between-subjects factor. The block of screening items always came last. We used this test structure because pilot work suggested that the order of the test questions could affect participants' accuracy and that blocking provided a more sensitive test of learning than intermingled question types.

2.2. Results

Preliminary analyses were run for each Trial Type to determine whether accuracy was influenced by the counterbalanced Languages (i.e., whether there was a difference in accuracy on the Non-adjacent items between participants in Languages 1A and 1B and participants in Languages 2A and 2B or a difference in accuracy on the Adjacent items between participants in Languages 1A and 2A and those in Languages 1B and 2B). There was no main effect of Language and no interactions between Language and the factors of interest. All subsequent analyses were therefore collapsed across Language.

2.2.1. Accuracy

Accuracy was quite good in all conditions for both Non-adjacent items and Adjacent items. The mean percentage correct for each Trial Type (Adjacent or Non-adjacent),

Exposure Length (2, 4, or 6 lists), and Test Order (Adjacent First or Non-adjacent First) is given in Table 1 and shown graphically in Fig. 2. Confidence intervals around each mean reveal that participants performed above chance (50%) at α of 0.05 for 9 of the 12 conditions (see Fig. 2). Participants demonstrated some learning of both dependencies after only 7 min of exposure (48 repetitions of each A_B frame) and more robust learning with more exposure.

To test the reliability of the differences between the means seen in Fig. 2, we fit the raw accuracy data with a logistic mixed-effects regression model using the lme4 package in R (Bates & Maechler, 2010; Jaeger, 2008; R Development Core Team, 2010). Trial Type (Non-adjacent vs. Adjacent), Exposure Length (2, 4, or 6 lists), and Test Order (Non-adjacent First vs. Adjacent First) were entered as fixed factors and Subject was entered as a random factor. The Trial Type and Test Order factors were coded for main effects. The model revealed a significant positive intercept, indicating that overall, participants were more likely to choose the correct than the incorrect answer ($b = 0.492$, $z = 8.247$, $p < .001$). There was no main effect of Trial Type, suggesting that accuracy was similar overall for Non-adjacent and Adjacent items. However, relative accuracy for Non-adjacent and Adjacent items varied by test order, as indicated by a significant two-

Table 1
Mean accuracy and confidence ratings (*SD*) for each condition in Experiment 1

| Test Order	Exposure Length	Accuracy		Confidence Rating	
		Adjacent	Non-Adjacent	Adjacent	Non-Adjacent
Adjacent First	2 lists	0.583 (0.227)	0.545 (0.214)	4.42 (1.16)	4.35 (1.20)
	4 lists	0.633 (0.225)	0.564 (0.267)	4.64 (0.88)	4.61 (1.27)
	6 lists	0.692 (0.184)	0.622 (0.247)	4.54 (0.94)	4.79 (1.08)
Non-adjacent First	2 lists	0.532 (0.231)	0.636 (0.257)	4.56 (0.96)	5.03 (1.20)
	4 lists	0.647 (0.178)	0.640 (0.281)	4.29 (1.00)	4.62 (1.28)
	6 lists	0.641 (0.181)	0.641 (0.278)	4.61 (0.95)	5.34 (1.37)
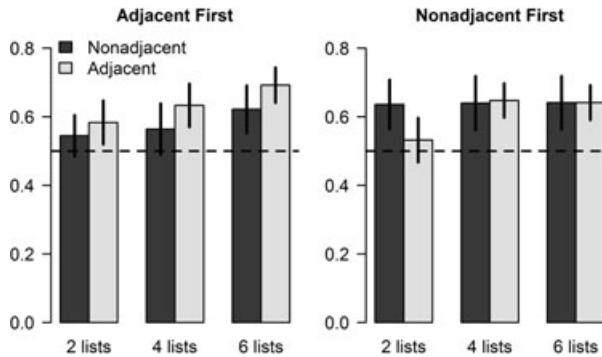
Fig. 2. Experiment 1 mean accuracy by Trial Type and Exposure Length for Adjacent First (left) and Non-adjacent First (right) participants. Error bars are 95% confidence intervals. The dotted line represents chance performance (50%).

way interaction between Trial Type and Test Order ($b = 0.407$, $z = 2.111$, $p = .035$). Accuracy improved with increasing exposure length ($b = 0.165$, $z = 2.267$, $p = .023$). No other main effects or interactions were significant.

To better understand the Trial Type × Test Order interaction, models were fit to each test order individually, with Trial Type and Exposure Length as fixed factors and Subject as a random factor. As expected from the pattern of means, model results for the two test orders were somewhat different. For participants tested first on Non-adjacent items, overall above-chance performance was supported with a positive intercept ($b = 0.541$, $z = 5.809$, $p < .001$). There were no other main effects or interactions (all $p > .15$). These participants performed equally well on Non-adjacent and Adjacent test items and performed about equally well at all exposure lengths.

Participants tested first on Adjacent items also performed above-chance overall, as indicated by the positive intercept ($b = 0.447$, $z = 5.877$, $p < .001$). There was a main effect of Trial Type ($b = 0.255$, $z = 1.888$, $p = .059$), with accuracy higher on Adjacent than Non-adjacent items. Participants were also more accurate following longer exposures ($b = 0.202$, $z = 2.160$, $p = .031$). The interaction was not significant, suggesting that accuracy increased for both Adjacent and Non-adjacent items as the exposure was lengthened.

This pattern of results demonstrates that participants were able to rapidly learn both structures, with some participants benefiting from longer exposure durations. The different outcomes obtained for the two test orders suggests that learning of the two structures was not equally robust. Participants did not simply score higher on the first block of test items than the second. Rather, participants tested first on Adjacent dependencies were subsequently less accurate in recognizing the Non-adjacent test items, but not the reverse. We will return to this point in the Discussion below.

2.2.2. Individual differences in learning

These analyses provide evidence of learning for both types of structure at the group level. While overall accuracy was approximately the same for Non-adjacent and Adjacent test items, inspection of individual scores gives us more information about the distribution of scores around the mean. The total number of items correct for each trial type for each participant is plotted as a frequency histogram in Fig. 3. Comparison of the two trial types reveals that while the means are similar, the distributions for the scores are quite different. Performance on the Non-adjacent items followed a bimodal distribution, with many participants scoring at chance (three items correct) and a significant minority getting all six items correct. However, performance on the Adjacent items followed a unimodal distribution centered on four items correct. This difference in distributions suggests that the Non-adjacent items were relatively easy for some participants—they were able to get all of them correct—while very few participants were able to answer as consistently on the Adjacent items. Less consistency on Adjacent items compared to Non-adjacent items makes sense given the structure of the stimuli. The adjacent structure consisted of 16 X words that occurred with different frequencies across frames, so that there were many items to track during training, and only a subset of the $AX_{HP}B$ items were represented in the test items.

In contrast, the non-adjacent structure consisted of only three deterministic word pairs, so that there was relatively less to learn and greater repetition during training, and all of the frames were represented in the test items. The analyses above suggested that Test Order influenced participants' relative accuracy on each item type. Consistent with this, 19 of the 28 participants who got perfect scores on the Non-adjacent items were in the Non-adjacent First condition.

While the overall means suggest that learners as a group were sensitive to both the non-adjacent and adjacent regularities, they do not indicate whether individual learners tracked both types of structure. If attentional demands require a trade-off between tracking
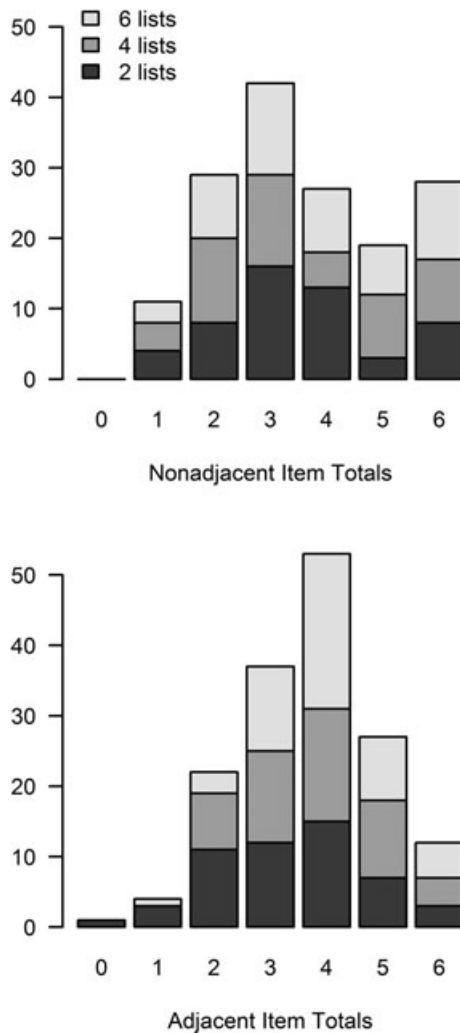

Fig. 3. Frequency histogram of accuracy scores for Experiment 1 for Non-adjacent items (top) and Adjacent items (bottom).

Table 2
The number of participants who were classified as learning each structure for Experiment 1

Exposure Length	Both Structures	Non-Adjacent Only	Adjacent Only	Neither
2 lists	10	15	15	12
4 lists	16	7	15	14
6 lists	20	7	18	7

adjacent patterns versus non-adjacent patterns, then we would expect most participants to show learning of one structure but not the other. If, on the other hand, the two types of structure reinforce one another, we would expect learners to master both. Our test design does not provide enough power to statistically test each learner's performance on each trial type. Instead, we scored each participant as a "learner" of a particular structure (adjacent or non-adjacent) if he or she scored above 50% on the test items for that structure and a "non-learner" if he or she scored 50% or below. In Table 2, we list how many participants were classified as learning each of the structures for each exposure length.

The fact that a substantial number of participants at all exposure lengths showed learning of both structures indicates that learners are capable of tracking both types of structure, even with relatively little experience with the language. With longer exposure, more participants learned both structures, fewer learned neither structure, and fewer learned only the non-adjacent structure compared with the shortest exposure. The difference in distribution of learners between the shortest and longest exposures is confirmed by a significant Chi-square test on rows 1 and 3 of Table 2 ($\chi^2(3) = 7.831$, $p = .050$). Interestingly, at the longer exposures, far more participants learned only the adjacent structure than only the non-adjacent structure. This asymmetry is partially due to the effect of Test Order discussed above. Participants who were tested first on the Adjacent items tended to score better on those items than the Non-adjacent items, a pattern we observed across exposure durations, whereas for the longer two exposures those participants tested first on the Non-adjacent items scored equally well on both item types.

It is also possible that differences in individual learning trajectories contributed to the changing distribution of learners. For example, it is possible that participants who learned the non-adjacent structure were likely to subsequently learn the adjacent structure, perhaps because knowing the non-adjacent structure allowed them to more easily track which X words were likely to be heard in each A_B frame (such as for frequent frames). Participants who focused exclusively on adjacent regularities, on the other hand, may successfully learn the adjacent structure, but may not attend to the non-adjacent frames enough to learn their structure. These possibilities will be addressed further in the discussion.

2.2.3. Ratings

The confidence ratings provide information concerning participants' awareness of their knowledge of the different types of structure during the test, and whether the different types of structure were equally salient to learners. Overall, participants rated themselves more confident on Non-adjacent test items than Adjacent test items, though this varied somewhat between conditions (see Table 1). To test this difference statistically, we fit the

ratings with a mixed effects model. The lme4 package does not provide p-values for the t-statistics (for discussion, see Baayen, Davidson, & Bates 2008), so the significance of interaction terms was determined using a log-likelihood ratio test between models including and excluding the interaction terms. An initial analysis including Trial Type (Non-adjacent vs. Adjacent), Test Order (Non-adjacent First vs. Adjacent First), and Exposure Length (2, 4, or 6 lists) as fixed factors and Subject as a random factor revealed no main effects or interactions involving Exposure Length (all $t < 2$). To simplify the model, Exposure Length was therefore removed as a factor (this did not change other effects). The subsequent analysis revealed a significant main effect of Trial Type ($b = 0.27$, $t = 4.71$) and a Trial Type × Test Order interaction ($b = 0.48$, $SE = 0.118$, $\chi^2(1) = 16.19$, $p < .001$); no other effects or interactions were significant. Participants tested first on the Non-adjacent items rated their confidence more highly for the Non-adjacent items than the Adjacent items, but participants tested first on Adjacent items rated their confidence the same for both trial types.

If confidence ratings are an accurate gauge of participants' knowledge, those participants who get more questions correct should rate their confidence higher than those who get fewer correct. Because participants varied in their absolute ratings, confidence ratings were normalized by subtracting the individual trial ratings from the participant's overall mean confidence rating, and the subsequent difference scores were correlated with participants' accuracy scores for each trial type. This analysis revealed significant positive correlations between rating and accuracy for all exposure lengths on the Non-adjacent test items (2 lists: $r = .481$, 4 lists: $r = .479$, 6 lists: $r = .471$, all $p < .001$) but not for the Adjacent test items (2 lists: $r = .025$, 4 lists: $r = .165$, 6 lists: $r = -.049$, all $p > .20$). Thus, participants who were more accurate on Non-adjacent items were also more confident in their ratings, while those that were more accurate on Adjacent items were no more confident than those who were less accurate. This difference in the relationship between ratings and accuracy for the two types of structure suggests that participants were more aware of their knowledge of the non-adjacent than the adjacent structure, even though they learned both structures similarly well.

2.3. Discussion

Participants in Experiment 1 demonstrated rapid learning of both non-adjacent and adjacent dependencies, with similar levels of accuracy on Non-adjacent and Adjacent test items. However, participants were more aware of the non-adjacent structure than the adjacent structure—they rated themselves more confident on Non-adjacent test items than Adjacent test items, and confidence ratings were correlated with accuracy only for Non-adjacent items.

Participants did not require long exposures to the language to learn the non-adjacent structure, as might be the case if participants initially default to tracking adjacent structure and only shift attention to non-adjacent structure when adjacent conditional probabilities are uninformative. In fact, the evidence points to participants tracking the non-adjacent structure early in exposure. After only a few minutes of experience with the

language, participants who were tested immediately on the non-adjacent structure performed above chance. Interestingly, increasing the exposure time did not have any effect on confidence ratings and only affected accuracy for some conditions. In particular, participants who were tested first on Adjacent items scored better on both Adjacent and Non-adjacent items with increasing exposure.

It is clear that the test order (whether Adjacent or Non-adjacent items were tested first) influenced participants' performance—this is expected given how different the familiarization and testing procedures were. The familiarization phase consisted of passively listening to the language while in the test phase participants actively selected items that sounded more like they belonged to the language they just heard. This selection process invites comparison of the test tokens both in our forced choice design and in test designs that use individual-item endorsement (e.g., Gómez, 2002). Pilot testing with interleaved trials resulted in poorer performance overall, so we adopted blocked test trials as a more sensitive design to detect learning. While participants may learn something about the artificial language's structure during familiarization, weaker representations may be disrupted by the test itself. Indeed, this appears to be the case in the current experiment, with accuracy on the non-adjacent structure being particularly susceptible to test order effects. At the two longer exposure lengths, participants were equally accurate on both types of structure when tested first on the non-adjacent items but were more accurate on the adjacent than non-adjacent items when tested first on the adjacent items. This finding is surprising, since the non-adjacent structure was deterministic and involved less vocabulary, resulting in much higher token frequency than the adjacent structure. In addition, each item testing the adjacent statistics used legal A_B frames, so these participants had more examples of correct non-adjacent structure before being tested than those in the other test order. It appears that when participants focused their attention on the X words and adjacent statistics they tended to weaken their representation of the A_B frames.

Examination of the individual participant data reveals that individual learners are capable of tracking both structures concurrently, though only some participants appeared to do so. At each exposure length some participants learned both structures, others learned only the adjacent or non-adjacent relationships, and some did not appear to learn anything at all. Interestingly, while the group means were fairly stable across exposure lengths, the proportion of participants in each of these groups changed with increasing exposure to the language. The number of participants who learned only the non-adjacent structure declined with more exposure (as did the number who did not learn), the number of participants who learned only the adjacent structure was similar across exposures, and the number of participants who learned both structures increased with exposure.

This pattern of individual outcomes, coupled with the group result showing that increased exposure primarily benefited learning of the adjacent dependencies, suggests that individual participants followed multiple different learning trajectories. It seems that participants who learned the non-adjacent structure most likely did so early in exposure, and that some of those participants subsequently learned the adjacent structure, perhaps by switching their attention from the frames to the X words within the frames. For these participants, the deterministic frame structure may have helped highlight the adjacent

structure in a manner similar to that proposed for categorization using frequent frames (Chemla, Mintz, Bernal, & Christophe, 2009). In contrast, focusing on the adjacent structure did not appear to highlight the non-adjacent structure, either during familiarization or test, perhaps because of the larger number of X words than frames. However, because our design was cross-sectional, these proposed learning trajectories are hypothetical and will need to be tested using a within-subjects design.

Some researchers have suggested that learning from the speech stream involves both a slow statistical learning mechanism responsible for representing transitional probabilities between syllables (adjacent and non-adjacent) and a rapid rule-learning mechanism that enables generalization beyond the items actually heard (Endress & Bonatti, 2007; Peña, Bonatti, Nespor, & Mehler, 2002). Crucially, these authors state that the rule-learning mechanism only operates on segmented speech, while the statistical mechanism operates on both segmented and continuous speech. While the rapid acquisition of the deterministic non-adjacent dependencies by participants in Experiment 1 may at first seem consistent with the rule learning mechanism proposed by Endress and Bonatti, our materials were not designed to assess rule learning. Endress and Bonatti and Peña et al.'s tasks were designed to test whether participants acquired sublexical rules, and generalized them to new items. The materials in this study focused on distributional regularities across words, rather than within words; it is not immediately clear how their proposed mechanisms would handle our materials. In addition, the rule-learning mechanism they propose is specifically intended to handle generalization beyond the input, which we did not test. Instead, we asked whether participants detected both the deterministic non-adjacent structure and the probabilistic adjacent structure, a question that required interrogating the specific words and structures presented during familiarization.

The fact that a substantial number of participants learned both adjacent and non-adjacent structures suggests that participants did not simply attend to either the adjacent relationships or the non-adjacent relationships but were able to track them both simultaneously and/or switch between them flexibly. It was not the case, for these participants, that the much stronger non-adjacent relationship caused participants to stop tracking the weaker adjacent statistics or that the presence of informative adjacent structure prevented the deployment of attentional resources to the non-adjacent structure. This finding is consistent with the results from Mintz (2002), who demonstrated categorization of words based on their co-occurrence in particular A_B frames. The current results suggest that participants can identify both the correct A_B frames and which X words were most likely to appear in those frames. Learning both of these structures would be required for successful categorization of words based on frequent frames.

The confidence ratings illuminated further differences between participants' knowledge of the non-adjacent and adjacent structures, with participants rating themselves more confident on their answers to the Non-adjacent items than the Adjacent items. Interestingly, the difference in salience of the two structures did not lead to differences in learnability. It is not surprising that the non-adjacent structure was more salient to participants than the adjacent structure: The non-adjacent structure was deterministic, and each A_B frame was far more frequent than any particular AXB string. What is more striking is that participants'

accuracy was related to their confidence in their responses for items testing non-adjacent structure but not items testing adjacent structure. Participants were not able to accurately gauge their knowledge of the adjacent structure, even when that structure was tested in the first block of test items. This suggests that the two types of structure were processed or represented differently by participants. The source of this difference is not immediately clear, given than the structures differed in type (adjacent vs. non-adjacent) as well as by strength of predictability (probabilistic vs. deterministic) and size of vocabulary (sixteen words vs. three word pairs); we will return to this point in the General Discussion.

The results of Experiment 1 make clear that learners are able to track both adjacent and non-adjacent linguistic structures within a relatively short exposure. However, it is less clear how the two types of structure interact during the learning process. We designed the artificial language in Experiment 1 to highlight the adjacent structure. While each A_B frame was heard equally often and with the same number of X words, the distribution of X words was very uneven, with the AX and XB conditional probabilities for $AX_{HP}B$ string four times greater than those for the $AX_{LP}B$ strings. It is possible that by highlighting the adjacent structure, we held some participants back from learning the non-adjacent structure (or from demonstrating their learning during test). In particular, the $X_{HP}B$ transitional probability was .80,[1] and this very high local transitional probability may have made participants less likely to keep the A word in memory. Thus, weakening the adjacent dependencies may boost learning of the non-adjacent dependencies.

It is also possible that the highlighting of the adjacent structure was *necessary* for the rapid learning of the non-adjacent structure that we observed in Experiment 1. Simple recurrent networks, which have been used to model human sequence learning, have difficulty tracking non-adjacent dependencies across irrelevant intervening items (e.g., Cleeremans & McClelland, 1991). However, when at least some of the intervening items are locally relevant (as in our artificial language), simple recurrent networks successfully learn the non-adjacent structure (Cleeremans, Servan-Schreiber, & McClelland, 1989). If human learning of non-adjacent dependencies also benefits from the presence of local dependencies, then strong adjacent dependencies may aid the process of acquiring non-adjacent dependencies. With weaker transitional probabilities between adjacent words, participants may fail to learn the non-adjacent structure or require longer exposures to do so.

Finally, it is possible that the two layers of structure do not "interact" at all. Participants may attend to non-adjacent and adjacent structure separately, regardless of the relative strength of the dependencies. For example, if participants' rapid acquisition of the non-adjacent dependencies in Experiment 1 was due to a fast-acting rule-learning mechanism that extracts the deterministic information, we would not expect changes in the strength of the adjacent dependencies to affect learning of the non-adjacent dependencies.

In Experiment 2, we sought to test how the strength of the adjacent dependencies influenced learning of the multiple structures. We designed a new artificial language with the same non-adjacent structure as Experiment 1 but with a more subtle manipulation of the

adjacent statistics. Within each A_B frame, the frequency of the X_{HP} words was decreased so that the AX_{HP} and $X_{HP}B$ transitional probabilities were only twice those of the AX_{LP} and $X_{LP}B$ transitions, rather than four times. This reduced the $X_{HP}B$ transitional probability to .67 from .80 and additionally ensured that X_{HP}, X_{LP}, and X_{ED} words occurred with equal frequency across the corpus. If the non-adjacent and adjacent regularities are learned independently, then we would expect learning outcomes to be similar to Experiment 1: Participants should demonstrate learning of non-adjacent dependencies after a brief exposure and additional exposure should primarily benefit learning the adjacent dependencies. However, if the two layers of structure do interact during learning, we would expect a different pattern of results than we observed in Experiment 1: The weaker adjacent dependencies could either cause participants to fail to learn the non-adjacent dependencies (or learn them only after longer exposures) or to succeed in learning the non-adjacent dependencies more easily, showing higher accuracy on Non-adjacent test items than Experiment 1.

3. Experiment 2

3.1. Method

3.1.1. Participants

One hundred and fifty-six participants (26 in each condition) were included in the analysis. An additional 35 participants were tested but excluded from the analysis due to missing three or more of the Screening test items (14); failure to respond to multiple test items (5); and experimenter or technical error (16).

3.1.2. Materials

Four new counterbalanced languages were created using the AXB strings from Experiment 1. Token frequency was changed in order to create the probability structure depicted in Fig. 1 (right panel). The corpus for the languages consisted of one repetition of each of the $AX_{ED}B$ strings, two repetitions of each of the $AX_{HP}B$ strings, and one repetition of each of the $AX_{LP}B$ strings, for a total of 48 tokens. See the Appendix for the exact strings and frequencies used. The 156 participants were randomly assigned to one of the languages in approximately equal numbers (1A = 39, 1B = 39, 2A = 36, 2B = 42). Within the six conditions determined by Exposure Length and Test Order ($N = 26$ in each) there were at least six participants assigned to each language condition.

3.1.3. Procedure

The procedure was identical to Experiment 1. Exposure times were slightly different because of the smaller size of the corpus. Thus, the shortest exposure (2 lists, 96 tokens) was 5 min and the middle exposure (4 lists, 192 tokens) was 10 min. For the longest exposure, we opted to give participants more experience with the language while equating listening time with Experiment 1, in order to obtain as strong an effect as possible. For

the longest exposure duration, participants listened for 20 min (as in Experiment 1), which was eight times through the corpus (384 tokens). The test items and procedure were the same as Experiment 1.

3.2. Results

The preliminary analyses revealed no differences in performance between Languages; all subsequent analyses are collapsed across Language.

3.2.1. Accuracy

Once again, participants were able to learn both the non-adjacent and adjacent dependencies. The mean percentage correct for each Trial Type (Adjacent or Non-adjacent), Exposure Length (2, 4, or 8 lists), and Test Order (Adjacent First or Non-adjacent First) is presented in Table 3 and displayed graphically in Fig. 4. Inspection of the 95% confidence intervals reveals that means were significantly above-chance performance (50%) for 9 of the 12 conditions (see Fig. 4). Evidence of learning is inconsistent at the shortest exposure length but more robust at longer exposure lengths.

We fit the accuracy data with the same logistic mixed-effects model used in Experiment 1, with Trial Type (Non-adjacent vs. Adjacent), Exposure Length (2, 4, or 8 lists), and Test Order (Non-adjacent First vs. Adjacent First) as fixed factors and Subject as a random factor. Once again there was an overall positive intercept, indicating that participants were more likely to choose the correct than the incorrect answer ($b = 0.533$, $z = 5.764$, $p < .001$). Performance again improved with increased exposure duration ($b = 0.086$, $z = 1.948$, $p = .051$). A significant interaction between Trial Type and Test Order indicated that relative accuracy on Non-adjacent and Adjacent test items was influenced by test order ($b = 0.395$, $z = 2.078$, $p = .038$). There were no other significant main effects or interactions.

While the patterns from the full model are largely similar to what we observed in Experiment 1, follow-up analyses on the individual test orders reveal differences between the results of the two experiments. Participants tested first on the Non-adjacent items performed above chance overall (positive intercept, $b = 0.546$, $z = 4.363$, $p < .001$) but

Table 3
Mean accuracy and confidence ratings (*SD*) for each condition in Experiment 2

| Test Order	Exposure Length	Accuracy		Confidence Rating	
		Adjacent	Non-Adjacent	Adjacent	Non-Adjacent
Adjacent First	2 lists	0.577 (0.190)	0.474 (0.154)	4.21 (1.06)	4.26 (1.13)
	4 lists	0.581 (0.251)	0.660 (0.238)	4.42 (0.80)	5.06 (1.30)
	8 lists	0.626 (0.199)	0.596 (0.246)	4.45 (0.89)	4.45 (1.32)
Non-adjacent First	2 lists	0.564 (0.194)	0.583 (0.201)	4.41 (0.76)	4.86 (0.86)
	4 lists	0.574 (0.189)	0.629 (0.300)	4.26 (1.24)	4.96 (1.60)
	8 lists	0.551 (0.175)	0.700 (0.289)	3.80 (1.02)	5.08 (1.41)
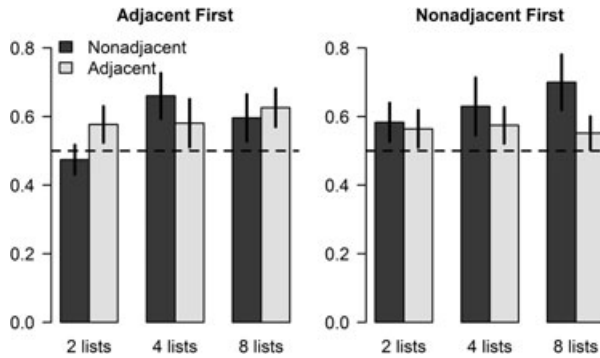
Fig. 4. Experiment 2 mean accuracy by Trial Type and Exposure Length for Adjacent First (left) and Non-adjacent First (right) participants. Error bars are 95% confidence intervals. The dotted line represents chance performance (50%).

also showed a significant effect of Trial Type, scoring higher on Non-adjacent than Adjacent items ($b = 0.648$, $z = 2.817$, $p = .005$); the interaction between Trial Type and Exposure Length was marginally significant ($b = 0.196$, $z = 1.800$, $p = .072$), providing some evidence that the effect of Trial Type was larger at longer Exposure Lengths. In Experiment 1, participants tested first on Non-adjacent items performed equally well on both item types and showed no effects of exposure length. With the less prominent adjacent statistics available in Experiment 2, however, participants' knowledge of the probabilistic adjacent structure was weak enough to be disrupted when they focused on the deterministic non-adjacent structure. Participants' knowledge of the non-adjacent structure also benefitted from more exposure to the language.

Participants tested first on Adjacent items also performed above chance overall (positive intercept, $b = 0.363$, $z = 4.378$, $p < .001$) but did not exhibit effects of Trial Type or of Exposure Length (all $p > .12$). This result is quite different from Experiment 1, in which Adjacent First participants were consistently more accurate for Adjacent items than Non-adjacent items but improved on both trial types with more exposure. This difference in performance is no doubt due to the greater challenge of learning the weaker adjacent statistics, particularly *relative* to the deterministic non-adjacent statistics. Importantly, though, participants were still able to learn the weaker adjacent dependencies, with above-chance performance at all three exposure durations.

3.2.2. Individual differences in learning

The distributions of individual scores for the two trial types reveal patterns of learning very consistent with Experiment 1, as shown in the histogram in Fig. 5. As in Experiment 1, the scores for the Non-adjacent items are bimodal, with many participants scoring at chance but a significant minority of participants getting all six items correct. The scores for the Adjacent items were unimodal and slightly lower than in Experiment 1. This pattern of results presumably reflects participants' greater difficulty in concurrently tracking the probabilistic adjacent dependencies and the deterministic

dependencies (especially relative to the more prominent adjacent dependencies in Experiment 1). As in Experiment 1, about two-thirds of the participants who got a perfect score on the Non-adjacent items (20 of 29) were in the Non-adjacent First condition.

Binning the participants as we did in Experiment 1, Table 4 shows the number of participants who learned each structure. A similar number of participants learned both structures as in Experiment 1, suggesting once again that some individuals were able to attend to and learn both the non-adjacent and adjacent regularities. With increasing exposure lengths, more participants learned both structures and fewer learned only one, changes
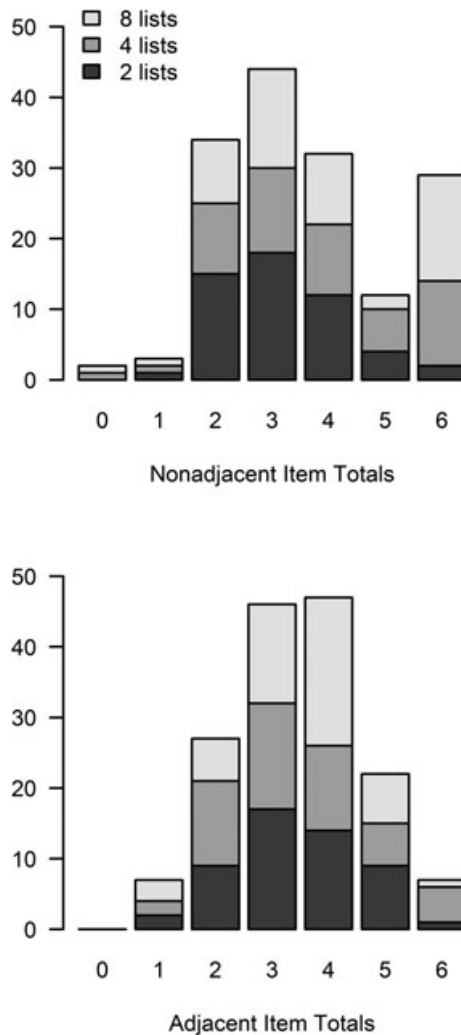

Fig. 5. Frequency histogram of accuracy scores for Experiment 2 for Non-adjacent items (top) and Adjacent items (bottom).

Table 4
The number of participants who were classified as learning each structure for Experiment 2

Exposure Length	Both Structures	Non-Adjacent Only	Adjacent Only	Neither
2 lists	5	13	19	15
4 lists	14	14	10	14
8 lists	18	11	11	12

that are reflected in a significant Chi-square test on rows 1 and 3 of Table 4 (χ^2(3) = 9.981, p = .019). Surprisingly, despite overall improvements in learning with more exposure, there is little decrease in the number of participants who learn neither structure, which remains at about 25% of the sample for all exposure lengths. Unlike Experiment 1, participants at the longest exposure who learned only one structure were evenly split between those who learned the non-adjacent structure and those who learned the adjacent structure. There is no evidence from the distribution of learners that participants first recognized the non-adjacent dependencies and then used them to help track the adjacent dependencies, as was suggested in the discussion of Experiment 1, but it is certainly possible that individual learners followed such a trajectory.

3.2.3. Ratings

As in Experiment 1, participants were more confident in their answers for the Non-adjacent than the Adjacent test items. As the patterns of means in Table 3 suggests, however, the effect of Trial Type was larger for participants who were tested first on the Non-adjacent items. We fit the ratings with a mixed-effects model with Trial Type (Non-adjacent vs. Adjacent), Test Order (Non-adjacent First vs. Adjacent First), and Exposure Length (2, 4, or 8 lists) as fixed factors and Subject as a random factor. This analysis revealed a significant three-way interaction (b = 0.167, SE = 0.049, χ^2(1) = 11.79 p < .001) as well as two-way interactions between Trial Type and Test Order (b = 0.475, t = 3.78) and Trial Type and Exposure Length (b = 0.054, t = 2.22) and a main effect of Trial Type (b = 0.486, t = 7.74). The Non-adjacent First participants consistently rated their confidence on Non-adjacent items more highly than the Adjacent items, and this difference increased with longer exposure. In contrast, the Adjacent First participants generated similar confidence ratings for the Non-adjacent and Adjacent items, except for the middle exposure length at which participants also rated the Non-adjacent items higher.

While the overall pattern of confidence ratings is similar to Experiment 1, one surprising difference is that participants' confidence on Adjacent items actually *decreased* as they gained more exposure to the language and Adjacent items were tested second. This suggests that given the lower adjacent transitional probabilities, participants' explicit awareness of the adjacent structure was relatively fragile and may actually have become less strong over time, at least relative to participants' awareness of the deterministic non-adjacent structure.

Further support for the hypothesis that knowledge of the non-adjacent dependencies is more explicit than that of the adjacent dependencies comes from inspecting the correlations between the normalized ratings and accuracy for each trial type. As in Experiment 1,

ratings and accuracy were correlated for Non-adjacent test items (2 lists: $r = .126$, $p = .370$; 4 lists: $r = .636$, $p < .001$; 8 lists: $r = .648$, $p < .001$), but not for Adjacent items (2 lists: $r = .067$; 4 lists: $r = .125$; 8 lists: $r = -.029$; all $p > .370$). This same result holds when the data are broken down by Test Order, with both Test Orders showing a positive correlation between Non-adjacent accuracy and ratings and neither showing a positive correlation between Adjacent accuracy and ratings.

3.3. Discussion

In Experiment 2, we replicated the findings from Experiment 1: Adults rapidly learn both adjacent and non-adjacent dependencies when both are present in an artificial language. Participants showed some learning of each structure at the shortest exposure length (96 tokens), though that knowledge was fragile and only detectable for the structure that was tested first. More robust learning was observed at the longer exposure lengths. While the mean accuracy was numerically greater for non-adjacent than adjacent test items at the longer exposures, this difference did not reach statistical significance. As in Experiment 1, confidence ratings revealed that participants' awareness of the non-adjacent structure was more robust than their awareness of the adjacent structure, and was a better predictor of accuracy for non-adjacent than adjacent structure.

Many findings were similar across experiments, including the overall accuracy for both Adjacent and Non-adjacent items. Weakening the adjacent statistics did not boost learning of the non-adjacent structure—participants did not attend to the non-adjacent structure to the exclusion of the adjacent structure. However, the weaker adjacent statistics also did not undermine learning of the non-adjacent structure, so it was not the case that strong adjacent statistics were *necessary* for participants to track the non-adjacent statistics. So do the two levels of structure interact at all? One difference between the two experiments lies in the pattern of accuracy for each test order. Relative accuracy for Adjacent and Non-adjacent items was influenced by test order in different ways across the two experiments, suggesting that the strength of the adjacent statistics affected participants' ability to maintain a representation of the non-adjacent structure. Relatedly, the pattern of individual results was somewhat different than what we observed in Experiment 1. Though a similar number of individual participants learned both structures at the longest exposure lengths across both experiments, the distribution of participants learning only one of the structures was different from that of Experiment 1, suggesting different learning outcomes across the two studies.

4. General discussion

Natural language contains many layers of distributional structure, and determining how humans concurrently track relationships across multiple structures will be crucial for understanding language acquisition, as well as learning in other richly structured domains. In two experiments we tested adult learning of multiple structures in an artificial

language. Experiment 1 demonstrated that adults were simultaneously sensitive to both adjacent and non-adjacent dependencies in the artificial language, though the deterministic (and more frequent) non-adjacent dependencies were more salient to learners. Experiment 1 also revealed significant differences in individual learning, with a large minority of participants learning both types of structure and smaller numbers of participants learning only one type of structure. Experiment 2 was designed to determine whether the strength of the adjacent dependencies influenced learning of both adjacent and non-adjacent dependencies. The results largely replicated those of Experiment 1, with participants showing sensitivity to both types of structure and the non-adjacent structure remaining more salient to participants. However, there were also differences in the outcomes of the two studies, particularly in how vulnerable learning was to disruption during test.

4.1. Strength of local dependencies affects learning

The materials used in Experiments 1 and 2 differed in the strength of the relationships between adjacent words (see Fig. 1). While the difference in absolute transitional probabilities between the two languages was not large, the gap between the higher and lower probability transitions was approximately twice as big for Experiment 1 than Experiment 2. Both languages supported learning of the non-adjacent dependencies, and at the longest exposure durations, a similar number of participants were able to learn both structures. However, closer inspection uncovers interesting differences between the results of the two experiments. One caveat to keep in mind when comparing effects across the experiments is that the corpus used in Experiment 1 was larger than the corpus used in Experiment 2, leading to a different number of tokens at each exposure level across the two experiments. Because of this, we discuss the differences between the experiments in qualitative terms but do not test them statistically.

Comparing the effects of Test Order across the two experiments reveals how the different statistical landscapes of the two languages influenced the strength of learning. It was not simply the case that participants had higher accuracy for the structure on which they were tested first. Rather, in each experiment one of the structures appeared more vulnerable to disruption during test than the other, as indicated by lower relative accuracy when tested second (but approximately equivalent accuracy when tested first). In Experiment 2, adjacent dependency learning was particularly susceptible to disruption. It is not surprising that it would be more difficult to maintain a representation of the adjacent dependencies than of the non-adjacent dependencies for the language used in Experiment 2. The non-adjacent dependencies consisted of the deterministic A_B frames, while the adjacent dependencies (both AX and XB) were much weaker. In Experiment 1, however, it was the deterministic non-adjacent structure that was more susceptible to disruption. The stronger adjacent statistics used in Experiment 1 appear to have made it challenging to maintain the non-adjacent structure while the adjacent structure was being tested. This result is especially surprising given that a number of factors in the design should actually have privileged the non-adjacent structure. As in Experiment 2, the non-adjacent dependencies were more predictable and involved less vocabulary than the adjacent

dependencies. Participants in Experiment 1 even had more exposure to the non-adjacent regularities than those in Experiment 2, due to the larger corpus used. Finally, even the test structure could have aided the non-adjacent dependencies: The items testing the adjacent structure all contained legal A_B frames, so that participants who were tested second on the non-adjacent structure had extra examples of that structure relative to those who were tested first. Despite these advantages, accuracy declined for participants tested second on the non-adjacent structure in Experiment 1.

Inspection of the individual data further supports the hypothesis that differences in the adjacent probability structure led to different learning outcomes in the two experiments. The individual data also suggest multiple possible learning trajectories. While a similar number of participants in both experiments were able to learn both types of structure, the number of participants who learned each individual structure differed across Experiments 1 and 2. To quantitatively test the presence of different patterns of learning across experiments, we combined the learner counts across the middle and long duration conditions of each experiment (for which performance was similar, see Tables 2 and 4). In Experiment 1, 14 participants learned only the non-adjacent structure and 33 learned only the adjacent structure, while in Experiment 2, 25 learned only the non-adjacent structure and 21 only the adjacent structure. A significant Chi-square test on these totals ($\chi^2(1) = 4.795$, $p = .029$) confirms the different patterns of learning across the two experiments. Given the weaker adjacent statistics in Experiment 2, more of the participants who learned just one level of structure apparently focused on the non-adjacent structure compared to Experiment 1.

The results of Experiments 1 and 2, taken together, suggest that the group data represent a heterogeneous mixture of learning trajectories that may map onto current theories of sequence learning. For example, some participants, particularly in Experiment 1, may have detected the non-adjacent structure and treated the strings similar to frequent frames— using the frames to track and categorize the X words. This seems particularly plausible given that our language, like the materials used by Mintz (2002), consisted of individual words that were arranged into phrases clearly delimited by longer pauses between phrases than between words. Other participants, particularly in Experiment 2, may have built up the sequence from adjacent elements, similar to the manner in which learning of sequential structure has been modeled using simple recurrent networks (Cleeremans & McClelland, 1991; Cleeremans et al., 1989). In this case, the adjacent statistics may have actually provided locally relevant context that facilitated learning of the non-adjacent statistics (e.g., participants learn that *pel* predicts *rud* by first noticing that *pel* predicts *vamey* and that *vamey* predicts *rud*). A more detailed investigation of individual learning trajectories is necessary to tease apart differences due to individual learners and differences due to language structure.

Another question of particular interest is whether learning of the different structures is truly simultaneous, resulting in incremental strengthening of representations throughout familiarization, or whether participants switch their attention between the two levels of structure, first learning some adjacent relations and then attending to the non-adjacent relations (or *vice versa*). Studies in which participants are either tested multiple times after multiple familiarization sessions or in which on-line learning measures are collected (such as

the AGL-SRT task developed by Misyak et al., 2010) could be used to reveal individual learning trajectories. One recent study did employ this paradigm to test learning of a language containing both adjacent and non-adjacent dependencies (Vuong, Meyer, & Christiansen, 2011). Consistent with the results of the current studies, they found that learners were sensitive to both types of dependencies from early in familiarization, though a long exposure duration was needed to reveal significant effects. Their design, however, did not provide a test of non-adjacency learning that controlled for adjacent statistics, limiting the conclusions that can be drawn, nor did they investigate individual learning trajectories.

The source of the individual differences in the current experiments remains unknown. However, recent work in other laboratories highlights the importance of understanding these differences. Statistical learning of both visual and auditory grammars is correlated with speech perception abilities in young adults, even when general cognitive abilities are controlled for (Conway et al., 2010). Additionally, learning non-adjacent dependencies in an artificial language is related to verbal working memory skills, while adjacent dependency learning is related to short-term memory span (Misyak & Christiansen, 2012). Our results suggest that adults vary not only in their ability to learn non-adjacent and adjacent dependencies but also in the process by which they learn such dependencies. It is possible that variation in verbal working memory and short-term memory skills influenced participants' learning trajectories in our task, though it is not clear whether the dissociation between dependency type and memory skill documented by Misyak and Christiansen would be obtained given materials like those used in the current experiments, in which both types of dependencies were embedded in the same language.

4.2. Different representation of non-adjacent and adjacent statistics

The fact that participants were able to learn both types of structure is particularly striking given the large difference in predictability between the adjacent and the non-adjacent structures. Participants neither focused on adjacent regularities to the exclusion of non-local dependencies nor attended only to the highly regular non-local dependencies to the exclusion of the much more variable adjacent relationships. While participants in both experiments were similarly accurate on items testing the adjacent and non-adjacent structures, they rated themselves more confident in their answers to the non-adjacent test items, particularly when they were tested first on the non-adjacent items. This is not altogether surprising; the deterministic structure and limited vocabulary of the non-adjacent structure may have made it more salient than the more nuanced adjacent structure.

More interesting than the difference in overall salience is the difference we observed in the relationship between participants' confidence and accuracy for the non-adjacent and adjacent structures. Confidence was correlated with accuracy only for non-adjacent test items, not adjacent test items. Participants were able to more accurately gauge their knowledge of the deterministic non-adjacent structure than the probabilistic adjacent structure, suggesting that the two types of regularities may have been represented differently during learning. Other researchers have found that confidence ratings are not always

related to accuracy (e.g., Dienes et al., 1995), and extensive research has investigated the relative implicit or explicit nature of sequence learning (for review see Cleeremans, Destrebecqz, & Boyer, 1998; Dienes & Perner, 1999), though this work has tended to use methodologies less directly related to language, such as serial reaction time or visual artificial grammar learning tasks (cf. Perruchet & Pacton, 2006). Because the two structures employed in the current studies varied in the their vocabulary size, distance, and reliability, we cannot say for certain whether the difference in knowledge awareness was driven by one or all of those factors. However, it is unlikely that learning of non-adjacent structures is *necessarily* explicit, given that implicit learning has been demonstrated in prior work investigating non-adjacent dependency learning. For example, Broadbent and colleagues investigated learning using a control task in which participants were instructed to generate a particular outcome from a computer program (Berry & Broadbent, 1988; Hayes & Broadbent, 1988). Results were consistent with explicit knowledge when the computer program's output on a particular trial depended on the participant's input on that trial, and implicit knowledge when output depended on the input on the previous trial. Importantly, the relationship between input and output was probabilistic in both conditions. Additionally, adults are able to learn phonotactic constraints across intervening consonants or vowels in procedures without explicit instruction (Newport & Aslin, 2004; Peña et al., 2002; Warker & Dell, 2006). Warker and Dell (2006) specifically compared a condition in which participants were told ahead of time about embedded regularities with one in which they were not told anything about the structure of the language. They found no difference in learning between the two groups.

It is also possible that our method of testing participants amplified differences between the two structure types. The alternative forced-choice procedure, which entails comparison of exemplars, may more easily detect explicit than implicit knowledge. If this is true, a procedure using a more implicit test of learning may be more sensitive for determining the relative strength of learning of the different structure types. On-line measures that do not distinguish between learning and test phases, such as serial reaction time (e.g., Misyak et al., 2010) or event-related potentials (e.g., Abla, Katahira, & Okanoya, 2008; Turk-Browne, Scholl, Chun, & Johnson, 2009), might be particularly useful in future work. Additionally, even our longest exposure duration was only about 20 min. The relatively short exposure times may have provided enough exemplars for participants to explicitly recognize the higher frequency non-adjacent structure but not the lower frequency adjacent structures. It is possible that longer exposure to the artificial language would lead to more explicit knowledge of the adjacent dependencies (i.e., correlations between accuracy and confidence ratings for both trial types).

The relative salience of different language-like structures and the relationship between salience and accuracy are important considerations for studies using artificial language learning as a model for natural language learning. For example, while knowledge of the non-adjacent structure was relatively explicit for the adult participants in our study, 15- to 18-month-old infants are also sensitive to non-adjacent structure in tasks using similar materials (Gómez, 2002; Gómez & Maye, 2005) and it seems unlikely that infants develop explicit representations during artificial language learning tasks.

Our data demonstrate that adults are capable of concurrently learning multiple levels of distributional structure. Similar results must be demonstrated in infants if this type of learning is relevant to first language acquisition. Infants have yet to be tested using materials containing both adjacent and non-adjacent dependencies, so it is unknown how learning such a language would compare with languages containing only one type of structure, or whether infants would be able to concurrently track both structures. The developmental trajectory of non-adjacent dependency learning has not been fully characterized, though it appears to improve over the second year of life: Gómez and Maye (2005) did not find evidence that 12-month olds could track the non-adjacent dependencies in their language; 15-month olds were able to do so, though perhaps less easily than 18-month olds. Prior experience with adjacent structure facilitates learning non-adjacent structure by 12-month-old infants (Lany & Gómez, 2008), suggesting that attention to non-adjacent statistics is influenced by context. While languages with multiple types of structures are inherently more complex than those with only one, recent work has demonstrated that infants are able to track adjacent transitional probabilities within a corpus of natural language (Pelucchi, Hay, & Saffran, 2009), suggesting that complexity does not necessarily overwhelm infants' statistical learning abilities.

Our studies provide initial data regarding the learning of multiple language structures. However, there are open questions that should be addressed by future work. In particular, while our data suggest that the non-adjacent and adjacent structure may have been processed differently, the source of that difference is unknown because of the difference in conditional probabilities in the two types of structure. Thus, we do not draw the conclusion that these dependencies must always be processed differently. The question of whether adjacent and non-adjacent dependency learning are subserved by the same or separate cognitive and neural systems is currently an area of intense investigation. At the neural level, which our data do not address, some authors propose that Broca's area is recruited specifically for hierarchically structured non-adjacent dependencies (e.g., Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006; Makuuchi, Bahlmann, Anwander, & Friederici, 2009), supporting the idea that adjacent and non-adjacent dependencies are necessarily processed differently. Others propose that Broca's area subserves processing of both local and long-distance dependencies (e.g., Petersson, Folia, & Hagoort, 2012).

At the cognitive level, some authors have proposed that the learning of non-adjacent dependencies is performed by a fast-acting rule-learning mechanism, while the learning of probabilistic adjacent dependencies is performed by a slow statistical learning system (Endress & Bonatti, 2007). Our data do not support this particular hypothesis, as they contradict several of the predictions for the rule learning mechanism. First, we found rapid learning of both adjacent and non-adjacent dependencies. Second, we found that participants' representations of the non-adjacent structure were strengthened by increased exposure to the language rather than being stable after an initial brief exposure. Finally, a rule-learning mechanism such as that proposed by Endress and Bonatti should be impervious to the adjacent statistics; the current results suggest that participants' learning of non-adjacent dependencies (identical across experiments) were affected by the strength of

the adjacent statistics (manipulated across experiments). However, our data do not rule out the possibility of multiple mechanisms for language learning.

5. Conclusion

Though we are modeling language learning using an artificial language, natural languages are far more complex. Our results suggest that complexity need not impede learning—indeed, multiple distributional structures may reinforce one another, even when the structures are not equally salient. Testing participants on multiple structures also revealed individual differences in learning, with some participants picking up on both structures and some only one (or none). In addition, the strength of the adjacent probability structure influenced learning outcomes and possibly the individual learning trajectories, suggesting that using multiple statistical structures within experiments may enrich our understanding of how participants learn sequential structure.

Acknowledgments

Note

1. The transitional probabilities given in the text are all *forward* transitional probabilities (e.g., $p(B|X_{HP}) = .80$. Because the non-adjacent dependencies are deterministic, the backward transitional probabilities mirror the forward probabilities (e.g., $p(A|X_{HP}) = .80$). Participants could draw on either forward or backward transitional probabilities (or both) to learn the adjacent structure.

References

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, *20*, 952–964.

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From learning items to generalizing rules. *Current Directions in Psychological Science*, *21*, 170–176.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-37. Available at http://CRAN.R-project.org/package=lme4. Accessed March 2, 2012.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, *79*, 251–272.

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*, 396–406.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*, 406–416.

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

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356–371.

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Learning, Memory*, *21*, 1322–1338.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *The Behavioral and Brain Sciences*, *22*, 735–755.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*, 247–299.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences*, *103*, 2458–2463.

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

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*, 183–206.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition*, *28*, 249–276.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

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

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*, 284–291.

Makuuchi, M., Bahlmann, J., Anwander, A., & Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences*, *106*, 8361–8367.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30*, 678–686.

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

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek, & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 31–63). New York: Oxford University Press.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302–331.

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.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*, 674–685.

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

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238.

Petersson, K. M., Folia, V., & Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain & Language*, *120*, 83–95.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 906–914.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*, 172–196.

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools, Inc..

St. Clair, M.C., Monaghan, P., & Christiansen, M.H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*, 341–360.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). "Frequent frames" in German child-directed speech: A limited cue to grammatical categories. *Cognitive Science*, *35*, 1190–1205.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*, 1–42.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*, 1934–1945.

Vuong, L. C., Meyer, A. S., & Christiansen, M. H. (2011). Simultaneous online tracking of adjacent and non-adjacent dependencies in statistical learning. In L. Carlson, C. Hölshcer, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 964–969). Austin, TX: Cognitive Science Society.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 387–398.

Appendix :

Languages 1A and 2A

A_B frame 1A	A_B frame 2A	X words	Relative Frequency Experiment 1	Relative Frequency Experiment 2
dak_tood	dak_rud	deecha, fengle, plizet, suleb	0	0
		balip, gensim, puser, vamey	1	1
		benez, kicey, loga, malsig	4	2
		gople, hiftam, roosa, skiger	1	1
pel_rud	pel_jic	deecha, fengle, plizet, suleb	1	1
		balip, gensim, puser, vamey	4	2
		benez, kicey, loga, malsig	0	0
		gople, hiftam, roosa, skiger	1	1
vot_jic	vot_tood	deecha, fengle, plizet, suleb	4	2
		balip, gensim, puser, vamey	0	0
		benez, kicey, loga, malsig	1	1
		gople, hiftam, roosa, skiger	1	1

Languages 1B and 2B

A_B frame 1B	A_B frame 2B	X words	Relative Frequency Experiment 1	Relative Frequency Experiment 2
dak_tood	dak_rud	benez, loga, plizet, roosa	0	0
		gensim, gople, skiger, vamey	1	1
		deecha, fengle, kicey, malsig	4	2
		balip, hiftam, puser, suleb	1	1
pel_rud	pel_jic	benez, loga, plizet, roosa	4	2
		gensim, gople, skiger, vamey	0	0
		deecha, fengle, kicey, malsig	1	1
		balip, hiftam, puser, suleb	1	1
vot_jic	vot_tood	benez, loga, plizet, roosa	1	1
		gensim, gople, skiger, vamey	4	2
		deecha, fengle, kicey, malsig	0	0
		balip, hiftam, puser, suleb	1	1

Test items

Each set of items was presented in random order to each participant.

Non-adjacent items

Languages 1A and 2A

String 1	String 2
dak gople tood	pel skiger jic
pel roosa rud	vot hiftam tood
vot hiftam jic	dak gople rud
dak hiftam rud	vot roosa jic
pel roosa jic	dak skiger tood
vot skiger tood	pel gople rud

Languages 1B and 2B

String 1	String 2
dak balip tood	pel suleb jic
pel hiftam rud	vot puser tood
vot puser jic	dak balip rud
dak puser rud	vot hiftam jic
pel hiftam jic	dak suleb tood
vot suleb tood	pel balip rud

Adjacent items

Languages 1A and 1B

String 1	String 2
dak benez tood	vot vamey jic
pel vamey rud	dak deecha tood
vot fengle jic	pel benez rud
pel loga rud	vot deecha jic
dak fengle tood	pel gensim rud
vot gensim jic	dak loga tood

Languages 2A and 2B

String 1	String 2
dak benez rud	vot vamey tood
pel vamey jic	dak deecha rud
vot fengle tood	pel benez jic
pel loga jic	vot deecha tood
dak fengle rud	pel gensim jic
vot gensim tood	dak loga rud

Screening items

Each incorrect item contains one of six X words not heard at all in the familiarization: *chila, coomo, nilbo, taspu, wadim,* and *wiffle.* Some incorrect items also contain an illegal A_B frame.

Language 1A

String 1	String 2
vot hiftam jic	dak chila tood
dak gople tood	pel nilbo rud
pel roosa rud	vot taspu jic
vot wiffle tood	pel vamey rud
dak wadim rud	vot gensim jic
pel coomo jic	dak benez tood

Language 1B

String 1	String 2
vot puser jic	dak wadim tood
dak balip tood	pel coomo rud
pel balip rud	vot wiffle jic
dak chila rud	vot vamey jic
pel nilbo jic	dak fengle tood
vot taspu tood	pel loga rud

Language 2A

String 1	String 2
pel skiger jic	dak chila rud
vot hiftam tood	pel nilbo jic
dak gople rud	vot taspu tood
vot wiffle jic	dak benez rud
dak wadim tood	pel vamey jic
pel coomo rud	vot fengle tood

Language 2B

String 1	String 2
vot suleb tood	dak wadim rud
dak puser rud	pel coomo jic
pel hiftam jic	vot wiffle tood
dak chila tood	pel loga jic
pel nilbo rud	vot gensim tood
vot taspu jic	dak fengle rud