Long-distance expectation-based sequence learning

Dylan Bumford[a]

[a]*New York University*

1. Introduction

During language acquisition, infants are faced with the task of extracting complex regularities from linguistic input, with little explicit feedback or training. It is by now uncontroversial that adults, children, and infants (Saffran et al. 1996b,a) alike display sensitivity to statistical patterns in many types of sequential data, even with fairly brief exposure. At a higher level of abstraction, individuals of all ages are also sensitive to a variety of *grammatical* patterns in sequential data (Reber 1967, Marcus et al. 1999). Given this large body of research demonstrating the lifelong human capacity for implicit sequence learning, it is an important for cognitive science to establish exactly *how* individuals come to recognize and utilize sequential patterns.

Several studies have examined the developmental trajectory of statistical learning abilities. Saffran (2002) tested adults and children for the influence of predictive dependencies between adjacent items in an artificial grammar, and found that while adults outperformed children overall, the two groups showed the same sensitivity to statistical relationships in the data. Gómez (2002) found similar correspondences between adults and 18 month old infants; both groups learned to recognize statistical dependencies between separated vocabulary items when the set of potential interveners was large, but not when it was small. Following up on this result, Gómez and Maye (2005) ran the same experiment on 12 and 15 month olds, demonstrating that even with a highly variable set of intervening elements, infants at 12 months failed to notice the dependencies between nonadjacent vocabulary items, but by 15 months, they performed like adults. This age range corresponds to the time in which children begin to show sensitivity to grammatical dependencies between separated morphemes (e.g. ‘*is* quickly runn*ing*’ vs. ‘*can* quickly runn*ing*’) (Santelmann and Jusczyk 1998).

These studies show that even relatively challenging sequence learning abilities — like tracking statistics between nonadjacent items — are broadly similar between children and adults, and demonstrate the utility of investigating how statistical learning abilities emerge and correlate with natural language acquisition. However, artificial language studies have largely prescinded from questions concerning the learning mechanism itself, focusing instead on the *outcomes* of learning tasks.

One counterexample to this comes from (Romberg and Saffran 2012), who presented infants with sequences of spatially-arranged targets, and measured their eye movements in anticipation of upcoming stimuli. For half of the infants, all of the first eight targets were presented in a single location, e.g. in the same place on a screen to the left of the infant. For the other half, six of the first eight targets appeared in that location on the left, but the second and sixth targets appeared on a screen to the right. Then for infants in both groups, the ninth target was presented in that secondary location. For the first group, this was the first appearance of an element on the right screen; for the second, it was the third appearance of an element on the right screen. They found that this low-probability event had a larger influence on subsequent expectations for the infants with variable early exposure than it did for the infants with uniform exposure. That is, infants in the uniform condition continued to look to the left screen in anticipation of the tenth trial, as they had before the anomalous event on the right, but infants in the probabilistic condition were more likely to look to the unlikely right screen on the tenth trial than they had been before. The authors concluded that the infants with a probabilistic model of the target location were more sensitive to prediction error, and more willing to adapt their expectations in response to unlikely events.

By monitoring infants' expectations about where the next item would appear, Saffran and Romberg were able to track the differential effect that low-probability items had on the two experimental groups *throughout the learning*. These expectations, as revealed through eye movements, provide the first glimpse into how particular data shaped the infants' underlying language model in real time.

Quite apart from infant artificial language studies, expectations have come to play a major explanatory role in several theories of online language comprehension. These theories are motivated by a variety of robust processing effects, primarily from self-paced reading and eye-tracking paradigms. For instance, it is well-established that the more predictable a word is, the faster it is read (e.g., Ehrlich and Rayner 1981, Van Berkum et al. 2005, DeLong and Urbach 2005). Similarly, visual world studies have shown time and again that both linguistic and extralinguistic context can lead listeners toward one interpretation of a structurally ambiguous phrase, well before the phrase is even completed (e.g. Trueswell et al. 1994, Tanenhaus et al. 1995, Altmann and Kamide 1999).

Results like these strongly suggest that individuals use available contextual and linguistic information *preemptively* zero in on the expected linguistic signal. Hale (2001) offers a model of incremental sentence parsing, according to which listeners maintain at all times a probability distribution over the possible completions of a partially revealed sentence. Building on this model, Levy (2008) argues that processing boosts/costs should be understood in terms of the relative redistribution of probability mass that new words require. Words that are already highly expected will not require any update to the subsequent predictions; but unexpected will require re-ranking of the competitors for the ultimate unfolding production. Taking this further, Pickering and Garrod (2013) have recently argued that internal forward modeling (i.e. making guesses about upcoming events) provides the substrate for all of human communication, including comprehension, production, and fluent dialogue.

Online predictions also play a central role in many computational learning models (e.g. Rescorla and Wagner 1972, Elman 1990). Chang et al. (2006) describe a learning algorithm that takes every opportunity to make predictions about upcoming material based on context and its current model of the sequence structure. Incorrect guesses spur the learner to adjust its underlying model, incrementally narrowing the gap between the predictions and the input. Jaeger and Snider (2013) outlines a similar procedure intended to capture the dynamics by which conversants adapt their own language models to converge on a common way of talking. This, they argue, accounts for the correlation between a syntactic structure's unpredictability and its effectiveness in priming subsequent discourse (Fine and Jaeger 2013).

One important aspect of these expectation-based learning and processing models is that they model the listener as considering predictions not just about the next word in an utterance, but about the entire future of the sentence. If these models accurately reflect the strategy of human learners, then it may be that making *long-distance predictions* — and attending to the subsequent long-distance feedback — is essential for grammar acquisition, especially for the acquisition of dependencies between nonadjacent lexical classes.

To address the extent to which sequence learning, and in particular context free grammar learning, is facilitated by ambitious prediction, I conducted an artificial language learning experiment. Participants watched as one cartoon character told a story to another, in a language that only the two characters understood. They were invited to imagine themselves as the character hearing the story. Often, this character would anticipate what the storyteller was going to say, and interrupt as if finishing the next chunk of the storyteller's sentence. On these occasions, the participants, playing the role of the interrupter, were invited to predict the next segment of the sentence, on the basis of the sentences they had seen before.

Given this setup, participants were divided into four groups. Half of the participants were asked to predict just the next word in the sentence, at the interruption point. The other half were asked to predict the next three words. Further, half of the participants were exposed to a language with relatively simple grammatical patterns, and half to a language with more complicated structures, of the sort found in natural languages. If predicting well into the future is important for learning complex structural patterns, as is commonly assumed by prediction-based learning algorithms, then we should expect to see an interaction between language complexity and prediction length. Participants encouraged to make longer predictions should prove better language learners, when the language to be learned contains complex structural dependencies, but not when it does not.

This experiment follows the lead of Romberg and Saffran (2012) in tracking online predictions throughout learning. The technique has several advantages. First, participants' predictions provide a window into their internal model of the language. We can see not merely which sentences an individual will *recognize* as (comparatively) likely, given some training data, but which sentences an individual will *generate*, using whatever patterns they have abstracted from the data so far. Second, as participants' understanding of the language changes, so will their predictions. Taken together, these two points give rise to a new analytical possibility: for any particular window of analysis, say the middle third of the participants guesses, we can use their predictions as a sample of the language, as they understand it at that time. From those samples, we can reconstruct pictures of the most likely underlying grammars that would generate those predictions. So where standard artificial language learning paradigms offer a single snapshot of participants' ability to discriminate grammatical from ungrammatical sequences at the end of learning, this prediction paradigm offers a series of snapshots of participants' approximations of the correct grammar, as they change over the course of learning.

2. Theoretical Background

2.1. Artificial Language Learning

2.1.1. AGL and SRT Paradigms

Sequence learning experiments generally follow one of two experimental paradigms: Artificial Grammar Learning (AGL) or Serial Reaction Time (SRT). In a typical AGL experiment (e.g. Reber 1967), participants are exposed to sequences of letters, syllables, shapes, scenes, tones, timbres, etc. that unbeknownst to them were generated by a particular grammar. After exposure, participants are presented with further novel sequences that conform to the rules of the grammar and similar sequences that do not. Very often participants can distinguish the novel grammatical sequences from the similar but ungrammatical sequences at levels greater than chance, suggesting they have learned something of the underlying statistical or algebraic structure of the language.

In SRT studies (e.g. Nissen and Bullemer 1987), participants are instructed to respond to sequentially presented stimuli, usually by pressing a specific button for each stimulus item, or by clicking a corresponding location on a screen. As in AGL studies, the stimulus locations are typically generated by a single grammar (typically less complex than in the AGL paradigm, though see Misyak et al. 2010, de Vries et al. 2012). Over the course of the experiment, participants acquire some knowledge about the regularities of the input, and respond more quickly to grammatical/predictable sequences of stimuli than to ungrammatical/unpredictable sequences (for reviews, see Clegg et al. 1998, Abrahamse et al. 2010).

Chang et al. (2012) point out that the same brain regions that support language processing also appear to support AGL and SRT behaviors, including Broca's Area (De Vries et al. 2010: Opitz and Friederici (2004)). What's more, functional imaging has revealed similar patterns of activation in and around Broca's Area during violations of natural language syntax and that of an artificial grammar as part of an AGL task (Petersson and Hagoort 2012). Christiansen et al. (2010) observed impairments in AGL performance in agrammatic aphasics. From these and other similar results, Chang et al. conclude that AGL and SRT tasks depend on the same mechanisms that underlie general linguistic competence and language learning.

2.1.2. Language Complexity

The complexity of a language is usually characterized by the kinds of formal grammars that can generate it. Chomsky (1956) famously described a hierarchy of formal grammars in terms of their respective expressive capacities. For natural animal languages, the most important distinction that the Chomsky hierarchy makes is between what are called Finite State Automata and Context Free Grammars.

Finite State Automata are simple machines consisting of a finite set of states, including one or more predefined initial and final positions, and a finite set of transition rules that push the grammar along from one state to the next. Each of the state transitions is associated with a symbol, generally thought of as a vocabulary item in the grammar being modeled. Thus as the machine transitions from state to state, it emits (or consumes) an element in the lexicon of the grammar. The sentences of the language then correspond to the different paths the machine may take to get from the first of its states to the last.

Automata like this can generate (equivalently, recognize) many different patterns, including patterns with infinitely many specific tokens (for instance, all strings consisting of an a followed by any number of bs and then another a), and even many aspects of natural languages. Other natural language patterns, however, are beyond their reach. The most famous example is the set of sentences of the form $a^n b^n$, where x^n signifies n repetitions of the item x. In English, for example, any sentence of the form '[the wolf]^{n}ate^n' is grammatical ('the wolf ate', '[the wolf the wolf ate] ate', '[the wolf [the wolf the wolf ate] ate] ate', etc.). Patterns like these require the grammar one way or another to keep track of which words appeared in the first half of the sentence, so that they can be appropriately paired with words in the second half. These sorts of pairings are called long-distance dependencies, and they in part motivate a more expressive approach to natural language modeling.

The simplest class of grammars that can capture $a^n b^n$ dependencies are called Context Free Grammars. One way to define a CFG is to start with an FSA and supplement it with a certain kind of very limited memory store called a pushdown stack. As the CFG moves through its abstract states, emitting items along the way, it can additionally add or remove items from an internal list that it maintains. This provides just enough capacity to effectively match the as with their corresponding bs. But there is another more common way of representing CFGs that doesn't rely on the notion of a memory. Instead, states are represented as nodes in an abstract tree, and transition rules determine how each node can unfold into subtrees. In this manner, the grammar may start in the middle of a pattern, so to speak, and emit the left and right portions of the string simultaneously, before dropping into a new "center-embedded" subtree. From this perspective, what distinguishes a CFG from an FSA is the ability to recurse into subpatterns *between two lexical elements*.

The point for the study of naturally-occurring communication systems is that different syntactic patterns place different lower bounds on the types of machines that might execute them. As a result, establishing that human languages recognize and utilize center-embedding schemes, for example, also establishes that the human language processing mechanism should have at least the computational power of a context free grammar. Obversely, if non-human animals do not exploit such patterns, it suggests particular limitations in the way they organize and manipulate sequential data.

2.1.3. Complexity and Learnability

A question that many sequence learning experiments have sought to address is to what extent the formal complexity of a particular pattern or language influences its learnability. As mentioned in the previous section, human languages support dependencies that cannot be expressed by finite state grammars. Many researchers believe that naturally-occurring musical systems likewise exceed the finite-state capacity (e.g. Lerdahl and Jackendoff 1983, Rohrmeier 2011). Nevertheless, it remains an open question what the upper limit is on people's capacity to absorb purely syntactic patterns from mere exposure, without explicit instruction.

For simple finite state automata, as well as probabilistic automata with only first-order conditional dependencies (either backward or forward), countless AGL and SRT experiments have established that adults can distinguish grammatical from ungrammatical sequences after relatively brief exposure (for an overview, see Pothos 2007). More surprising, infants are likewise sensitive to low-order statistical regularity (e.g. Saffran et al. 1996a, Aslin et al. 1998, Maye et al. 2002), as are several primate and bird species (e.g. Fitch and Hauser 2004, Gentner 2007).

Non-adjacent dependencies present a different story. Some kinds of dependencies are susceptible to implicit learning, including token identity (Gomez et al. 2000, Marcus et al. 1999) and more generally perceptual similarity (Creel et al. 2004, Gebhart et al. 2009). However, arbitrary non-local associations between stimulus items, of the sort found in natural languages, have generally eluded attempts at implicit learning. Perruchet and Rey (2005), Friederici et al (XXXX), Hochmann et al. (2008), and de Vries et al. (2008) all exposed individuals to properly context-free language fragments, but participants did not successfully distinguish grammatical test sequences from similar but ungrammatical sequences.

However, Gómez (2002) discovered that simpler non-adjacent dependencies were learned when the intervening material was sufficiently variable. That is, when faced with a larger vocabulary and therefore a larger overall number of unpredictive first-order transitions, Gomez's participants began to track the much more predictable second-order associations. This suggests that greater statistical variety in the sequence data may actually encourage people to entertain more complex patterns that they would not otherwise bother with. Consonant with this, a few recent AGL studies have reported success with context-free and even context-sensitive patterns (Lai and Poletiek 2011, Uddén et al. 2012, de Vries et al. 2012, Rohrmeier et al. 2012), either by substantially increasing the duration of exposure (Lai and Poletiek 2011, Uddén et al. 2012), or by increasing the vocabulary, sequence length, and/or structural variety of the languages being learned (de Vries et al. 2012, Rohrmeier et al. 2012).

2.2. Expectation-Based Learning

One of the themes running through computational and cognitive neuroscience in the last quarter century is a view of the brain as a finely-tuned prediction machine, a "bundle of cells that supports perception and action by constantly attempting to match incoming sensory inputs with top-down expectations or predictions" (Clark 2013). On this view, the primary activity of the brain consists in using sensory data, statistical information, and abstract, representational, or generative models of the world to make predictions about what will happen next.

In line with this general trend in cognitive science, syntactic prediction error has been argued to play an important role in many models of language processing. Here, 'syntactic prediction error' is taken to refer to the deviation between what words are observed and what words were expected to be observed prior to the observation. A growing body of research indicates that online processing difficulty is correlated with this sort of prediction error (for recent presentations of the program, see Federmeier 2007, Kamide 2008, Kutas et al. 2011, Levy 2013).

For instance, Hale (2001) and following him Levy (2008) hypothesize that individuals use stochastic context free language models to assign probability distributions to sentences in real time. According to the model, at each word of an unfolding sentence, an individual will re-weight all of the possible syntactic structures that might have generated the observed words as an initial segment, assuming one of these (hopefully the most likely) will correspond to the actual sentence once completed. They argue that garden path effects should be understood as points in a sentence at which a word is particularly unexpected, and therefore forces a substantial re-distribution of probability mass over possible continuations.

In a separate, but related line of research, Hale (2006) proposed that processing time should be proportional not (only) to the surprisal at a particular word (essentially a measure of conditional probability), but to the contribution of a word in reducing uncertainty about the future of the sentence. That is, certain words will provide a great deal of information about the future of the sentence, shifting the nature of the implicit probability distribution individuals maintain over possible continuations.

The operative assumption in of these sorts of information-theoretic processing models is that both speakers and comprehenders are in the business of constantly surveying a large collection of potential sentences, and ranking them in terms of their compatibility with what has been understood so far. Some models calculate rankings in terms of rich probabilistic context free grammars (Jurafsky 1996, Hale 2006, Levy 2008, Linzen and Jaeger 2014); others in terms of part-of-speech sequences (Frank 2010, Blache and Rauzy 2011); still others direct lexical and collocational statistics (Frank 2013); or some combination of these (Roark et al. 2009, Wu et al. 2010). But one way or another, they all aim to explain sentence processing costs in terms of the projected future possibilities for the sentence.

Increasingly, researchers have also begun to emphasize the role of prediction and feedback in language acquisition and adaptation, in addition to online sentence processing. Trueswell et al. (2013) propose that children make very specific predictions about the referents of lexical items, and only revise those predictions upon evidence to the contrary. Similarly, Mani and Huettig (2012) identify correlations between children's language production ability and their ability to predict upcoming words in sentences they hear. Conway et al. (2010) found that individual differences in implicit learning abilities were correlated with how well an individual is able to use word predictability to guide language perception. Change et al. (2006; 2012), Pickering and Garrod (2013), and Jaeger and Snider (2013) all argue that syntactic abstraction and sentence production crucially depend on learners' predictions about upcoming words during comprehension. In fact, according to the expectation-adaptation model adopted by Fine and Jaeger (2013) and Jaeger and Snider (2013), even adult conversants use prediction errors to rapidly and continuously adjust their underlying language models, so as to converge on a shared probabilistic grammar that maximizes mutual syntactic predictability.

Despite the overwhelming evidence that listeners make incremental predictions about upcoming syntactic structure, and the emerging view that such predictions drive both short- and long-term implicit learning, there is very little experimental language learning research that engages directly with expectation-based learning. Misyak et al. (2010) designed an SRT task around triplets of nonce words, drawn from a grammar with probabilistic dependencies between the first and last words of the triplets. Before each trial, participants were presented with two word candidates for each position of the upcoming triplet, and instructed to click on each word that they heard as soon as they had heard it. Of course, participants could boost their reaction times by anticipating later words on the basis of earlier ones, and moving the mouse to the expected word in advance. They found that (some) participants' reaction times to words in the third position decreased over the course of the experiment, indicating that they had picked up on the non-adjacent dependencies.

Alexandre (2010) conducted an SRT experiment based on a properly context free "palindrome language" of stimulus locations (e.g. XYZZYX, XWYWWYWX, etc.). He compared the reaction times of participants to the surprisal values that would be predicted by various language models. He found that the reactions of participants were equally well approximated by an SRN and a PCFG trained on the same sequential data, though not very well approximated by n-gram models or Markovian models. Similarly, Gureckis and Love (2010) designed two different artificial languages, one well-suited to the learning mechanism of a Simple Recurrent Network and the other well-suited to that of a Linear Associative Shift-Register. They then trained the two machine learners on the two languages, and compared the models' prediction accuracies at various points in the input sequences with the response times of human participants. Reaction times were better modeled by the simple LASR learning model than the more powerful SRN.

Each of these recent studies approximated sequence item predictability with an online response time measure, and exploited the SRT paradigm's access to response data throughout the experiment to draw conclusions about how participants learned the respective languages. In effect, they presuppose that participants consistently anticipate upcoming stimuli, and then — from an approximate measure of those anticipations — attempt to reverse-engineer a learning algorithm that would predict the empirical pattern of anticipations.

The present investigation begins from the same assumptions about anticipation, and likewise seeks to record some measure of participants' trial-by-trial learning trajectories. But where the previous studies took expectations for granted, exploiting anticipatory and reactionary behavior as indirect measures of language mastery, this study investigates the role that expectations themselves play in the learning process. To this end, I ask whether expectations themselves can be manipulated to facilitate different sorts of grammar learning.

For instance, the information-theoretic processing models introduced above all require individuals to compute the probabilities of many potential *complete sentences* compatible with some observed partial sequence. Surprisal models, which identify the processing cost of a word with the statistical predictability of that word given its preceding linguistic context, implicitly marginalize over all possible *complete* structures that would yield that word at that location. Entropy-reduction models are more explicit about the vast degree of forward modeling they assume; to compute the relevant quantity of information provided by a word in context, one should know not just the probabilities of the next potential items, but the shape of the entire probability distribution over all possible completions of the sentence.

Linzen and Jaeger (2014) argue on the basis of self-paced reading time data that these entropy-reduction values based on deep, farsighted predictions model behavioral processing cost more accurately than simple conditional probabilities or shallow predictions based only on the probabilities of the next (several) items (cf. Roark et al. (2009) and Frank (2013)). So, on the one hand, prediction error is increasingly recognized to figure prominently in language learning and adaptation. On the other hand, emerging evidence from computational models of online sentence processing suggest that individuals may routinely anticipate a great deal of linguistic material, many words into the future of the sentence.

Taken together, this suggests a natural hypothesis for expectation-based models of language learning: mastering complex sequential patterns, especially those with natural-language-like long-distance dependencies, may depend on the ability of individuals to make long-distance predictions about downstream material. That is, given the importance of prediction error in sequential learning, and the importance of fine-grained prediction in real-time sentence comprehension, we might wonder about the interaction of these phenomena, i.e. whether deep downstream prediction plays an important role in language learning. I hypothesize that it does, and the role it plays is in providing feedback on potential linguistic patterns that span multiple intervening elements. More specifically, I hypothesize that predictions about faraway items facilitate learning the recursive context free patterns that characterize natural human languages.

If complex pattern learning requires long-distance prediction, then one explanation for the frequent failure of participants to notice long-distance associations in artificial language learning tasks (see discussion in §2.1.3) could be that they are simply not making predictions far enough into the future. As a result, (assuming that individuals only utilize the parts of the input that they make predictions about) they only receive data about adjacent transitions, and thus only formulate, test, and confirm hypotheses about local dependencies.

Rohrmeier et al. (2012) reason similarly. They consider the hypothesis that long-distance dependencies go undetected in laboratory environments because the grammatical patterns are unnaturally simple. Putting this to the test, they found that in a relatively standard AGL experiment based on more ecological context-free structures, participants did show implicit (though not explicit) sensitivity to long-distance dependencies in the test sentences. Likewise, Thompson and Newport (2007) created artificial languages with naturalistic syntactic properties, including optional phrases, repeated phrases, moved phrases, differentially-sized lexical categories, and combinations thereof. Discrimination ability improved across participants as the languages increased in complexity. Gómez (2002) too found that increased trigram variability facilitated non-adjacent dependency learning.

Framed in terms of prediction error, these results show that without sufficient complexity and variety in the language, participants only bother making predictions about local dependencies. The following experiment tests this hypothesis by encouraging participants to make explicit predictions of varying lengths while learning either a simple (finite state) or complex (context free) language.

To the best of my knowledge, this is the first experiment to attempt to directly manipulate prediction behavior, as a potential causal factor in sequence learning. Pitting expectations against learning in different grammatical environments brings theories of information-theoretic language processing head to head with theories of expectation-based language learning, which brings us closer to a robust understanding of how synchronic online performance models and diachronic developmental learning models are related.

In addition, eliciting predictions throughout the experiment offers a rich window into the development of language proficiency. As with SRT studies, this continuous measure provides some indication of the time course of learning in different experimental conditions.

But unlike the SRT paradigm, sentence predictions also provide a kind of *production* data. So we can say with some confidence not just *whether* a given sequence was likely to be recognized by a given participant at a given stage of the experiment, but *which* sequence a participant was likely to generate at a given stage.

Thus, regardless of the particular results to this preliminary investigation, the present design offers two sources of rich empirical data that should be of interest to the experimental language learning community: direct access to online expectations, and iterative sequential productions throughout the learning period.

3. Experiment

The goal of Experiment 1 was to determine whether the acquisition of complex linguistic patterns could be facilitated by encouraging individuals to make predictions about larger stretches of upcoming items. To this end, participants were presented with a number of sentences drawn from one of two synthesized languages. The first language was generated by a simple finite-state grammar with no recursion. The second language was generated by a context-free grammar with up to three levels of center-embedding recursion. After sufficient exposure, participants were asked to make predictions about parts of sentences that they hadn't yet seen. Some participants were asked to predict a single word; others three words. According to the reasoning of §2.2, those participants encouraged to predict longer chunks of the sequences thereby received more implicit feedback about the language, and crucially more feedback about the connections between non-adjacent components of the grammar.

As a result, an interaction was predicted between grammar complexity and prediction window length. For those participants dealing with the finite-state grammar, prediction length should not make a difference, since these language satisfy the first-order Markov assumption: $p(w_i \mid w_{i-2}, w_{i-1}) = p(w_i \mid w_{i-1})$. But for those participants dealing with the context-free grammar, prediction length matters, if the hypothesis considered above is correct. Longer prediction windows should encourage participants to formulate and test hypotheses about the structure of the language that exceed the expressive power of finite-state models, and crucially exceed the hypothesis space available to participants making predictions about first-order transition probabilities alone.

3.1. Method

3.1.1. Participants

122 participants were recruited via Amazon Mechanical Turk. Participants received $\$5.00$ dollars for participation. All participants provided informed consent prior to the experiment.

3.1.2. Stimuli and Materials

Participants were randomly assigned to one of four groups. Half of the participants were exposed to sequences generated by a finite-state grammar, and half to sequences generated by a recursive context-free grammar (conditions F and C, respectively). Half of each of those groups were asked to predict a single word in each sequence, and half were asked to predict three words (conditions 1 and 3, respectively).

3.1.3. The Grammars

The finite state language was based on the BROCANTO language used in Opitz and Friederici (2007). It is specified in Figure 1a. The language designates six categories of terminals, glossed here for convenience with familiar parts-of-speech labels. Three of the categories (P, D1, and D2) are associated with a single word, and so require no generalization. The other three categories (N, V, and A) were each equally likely to be realized as one of two possible words. Mastering the grammar requires identifying the equivalence of these words with respect to the sequential patterns in the data.

Lexical items from Thompson and Newport (2007): roughly equivalent with respect to meaningfulness (Archer 1960), and all in the mid-range of phonotactic neighborhood density (Vitevitch and Luce 1999), i.e. all medium-probable nonwords.

The context free language was based on the left-branching, center-embedding grammar of Rohrmeier et al. (2012). It is specified in Figure 1b. Again, the language designates six categories of terminals, three of which comprise a single word, and three of which comprise two equally probable words.

S	→	NP TP
TP	→	VP (PP)
VP	→	V (A) | V (A) NP
NP	→	D1 N | D2 A N
PP	→	P NP
N	→	"daz" | "mer"
V	→	"lev" | "jes"
A	→	"tid" | "rud"
P	→	"nav"
D1	→	"sib"
D2	→	"zor"

(a) Finite state grammar based on Opitz and Friederici 2007

S	→	NP VP
VP	→	V1 | V2 NP
NP	→	N D1 | CP N D2
CP	→	VP R
V1	→	"daz" | "mer"
V2	→	"lev" | "jes"
N	→	"tid" | "rud"
R	→	"nav"
D1	→	"sib"
D2	→	"zor"

(b) Context free grammar based on Rohrmeier et al. 2012

Figure 1: Participants were exposed to, and made predictions regarding, sentences in one of two artificial languages. **(a)** Half of the participants saw sentences generated by a finite state grammar with no long-distance dependencies. **(b)** The other half saw sentences generated by a context free grammar with multiple layers of potential recursive embedding.

The two languages have the same nine-word vocabulary, the same number of lexical categories, and essentially the same number of production rules. Both have been shown to be (implicitly) learnable in standard artificial grammar paradigms (Opitz and Friederici 2007: Rohrmeier et al. (2012)). The primary difference between them is that the former is finite, and the latter infinite as a result of its recursive NP → CP → VP → NP productions.

3.1.4. Stimuli

Rohrmeier et al. exhaustively enumerated all sentences up to three levels of embedding, and then randomly selected 25 1-layer structures, 90 2-layer structures, and 81 3-layer structures, for a total of 195 different sequences.[1] The CFG stimuli used in the present study are composed of these same 195 sequences, modified slightly to include terminals representing the D1 and D2 categories, which were not present in Rohrmeier's et al. language.

For each length n sequence of the CFG language adapted from Rohrmeier et al., a length n sequence was randomly selected from the FSA language in Figure 1a. Thus the stimuli in the two conditions consisted of the same number of length n sentences, for $3 \leq n \leq 11$, and thus the same number of sentences overall.

3.1.5. Stimulus Rendering

The entire experiment was conducted over the internet, in a browser window of the participant's choice. Each trial consisted of a single sentence, revealed revealed one word at a time, from left to right, as shown in Figure 1b. Trials were separated by 1 second of inactivity. During the presentation of each sentence, a new word was revealed every 250 ms, until the prediction point, if there was one, or the end of the sentence otherwise. When predicting, participants selected either one or three words from a list of the available vocabulary items. No time limits were placed on prediction.

3.1.6. Procedure

Participants were informed that they would participate in a discourse between two cartoon characters. In the experimental conceit, one of the two characters was to tell the other one (played by the participant) a story. The story would be in a language the characters understood (but the participant obviously did not). At various points, the participant's character would anticipate the storyteller's next words, and interrupt the storyteller in excitement, much the way that people sometimes finish each other's sentences when they know what's coming.

Given this setup, participants first witnessed the storyteller utter 34 sequences, each between 3 and 5 syllables in length, to get some feel for the basic constituents of the language. Lai and Poletiek (2011) found that staging the input in this fashion — structures with no or limited recursive embedding before structures with multiple layers of hierarchical dependency — improved implicit acquisition of context free languages.

After this initial exposure to the language, participants transitioned into the second phase of the experiment, in which they were required to make predictions about various segments of the sentences they saw. On a prediction trial, participants saw an initial vertical red bar at the point in the sentence where the "interruption" was to occur (see Figure 2). If the sentence was 9 words long, for example, the interruption point might occur after the fourth word. At this point, partic-

[1] These figures combine the sentences generated for training purposes with those generated for testing

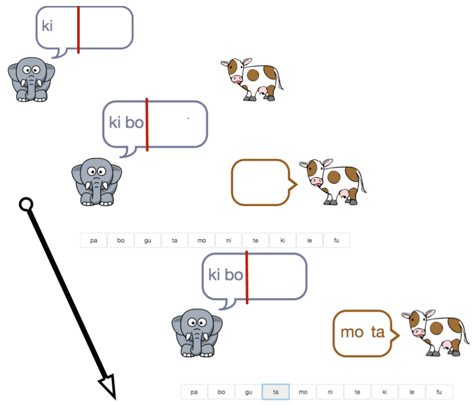

Figure 2: Schematic of a trial dynamic. At the beginning of every trial, a speech bubble appeared above the elephant, together with a line indicating the upcoming point of "interruption". One word at a time would appear in the elephant's speech bubble until the line was met, at which point a speech bubble appeared next to the cow, indicating the participant was to make a prediction about the next word (or words). The participant clicked on the buttons below the characters to make their selection(s). After they guessed, the elephant would react, and then finish its own sentence.

ipants would have to select whichever word or words (depending on the condition) they expected to come next in the sentence.

Once the participant made his or her selection, the storyteller reacted either positively or negatively, to indicate whether the prediction was correct or incorrect. Only exact word-for-word matches were considered correct for the purposes of feedback, though in fact on most trials, several of the choices could have led to "grammatical" sentences. For instance, given the finite state language in Figure 1a, there are two grammatical candidates for the first missing word after 'tid' in Figure 3, corresponding to the two parses shown below. The unknown material could constitute a PP, in which case the missing word would be 'nav', or it could constitute a NP, in which the missing word would be 'zor'. In addition to this structural uncertainty, participants also had to deal with lexical uncertainty. That is, even after correctly guessing that 'nav' should follow 'tid' (and therefore that the following word should be 'sib'), the final word could be either 'daz' or 'mer', as both are perfectly acceptable Ns in that syntactic context. For the purposes of scoring the participants answers, any word of the appropriate lexical category was counted as correct.

No matter what the participant guessed, after reacting, the elephant finished its sentence so that participants saw the intended words, as well as the remainder of the sentence that followed the material they were asked to predict.

This process was repeated until the story was complete, i.e. until all 195 stimulus sentences had been seen. Altogether, each subject provided 165 predictions, one for each sentence length 6 or longer and a quarter of those of length 5.

3.2. Results

3.2.1. Data Quality

Four of 122 participants, one from each condition, were recorded as answering correctly on more than 90% of the 165 trials. Given the indeterminacy inherent in the task, even with perfect knowledge of the grammar from the very beginning, performance at this level is all but impossible. These participants were excluded from the analysis. All other participants showed overall accuracies within three standard deviations of the mean, and no other measure was taken to exclude data. This left 118 participants.

Each of these participants completed 165 trials, some predicting three words per trial, others predicting just one. To maintain evaluation parity between the conditions, only the first prediction of each trial was scored. As explained in the preceding section, if the participant selected a word of the same category as the actual word in the sequence, their prediction was counted as correct. Otherwise, it was counted as incorrect. For each of the 118 participants, all 165 predictions were included in the analysis.

3.2.2. Final Performance

I hypothesized that a longer prediction window would facilitate context-free learning to a greater extent than it would finite-state learning. To test this hypothesis, I first restricted attention to a final subset of trials, after some amount of learning could be seen to take place. In this design, stimuli were not divided into separate training and test sets, and performance was assessed continuously throughout the experiment. Thus, it was planned in advance to treat the first 100 trials as a period of learning, leaving the last 40% of the stimuli for assessment.

The accuracies of participants in this later stage of the experiment are shown in Figure 4. On each trial,

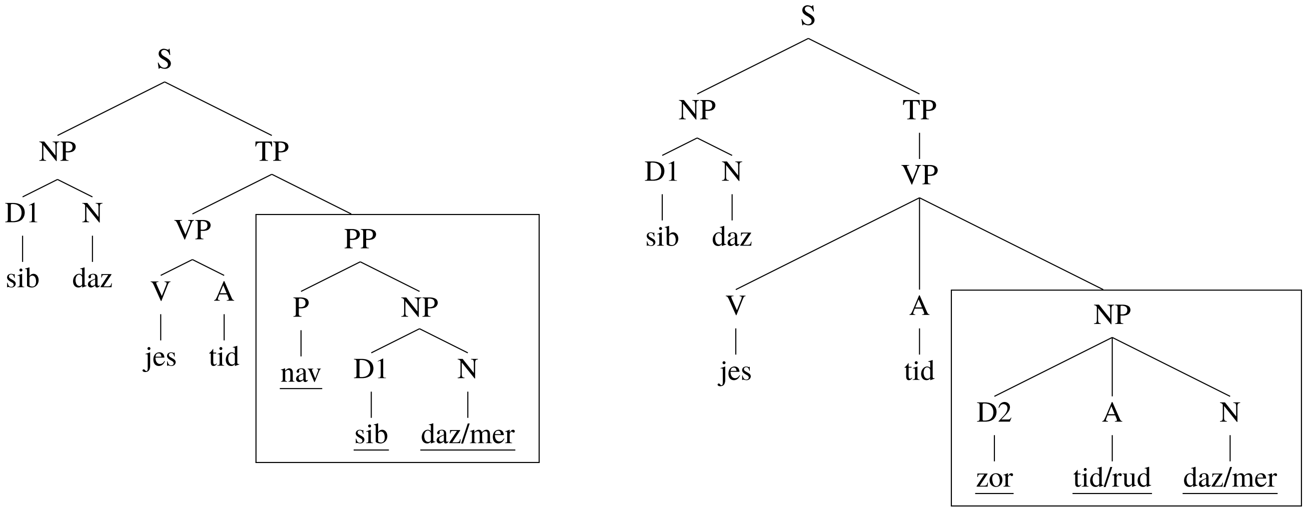

Figure 3: Possible continuations of 'sib daz jes tid X X X', according to the finite state grammar in Figure 1a. Participants were counted as correct if they guessed any word of the appropriate category. For instance, if the actual sequence was 'sib daz jes tid nav sib daz', credit was assigned for either 'daz' or 'mer' in the final slot.

there were nine word choices available to participants, so a random guessing strategy provides a baseline of around 11%. Nearly all participants exceeded this baseline, and average accuracies in each of the four conditions were well above this chance level, ranging from 23.3% to 32.2%.

For both languages, participants asked to predict just the next word in the sentence outperformed those asked to predict the next three words. However, the magnitude of the difference varied between the languages. For those learning the finite state language, one-word prediction-window participants were 38% more likely to guess correctly than three-word prediction-window participants. But for those learning the context free language, one-word participants were only 9% more likely to guess correctly than three-word participants.

To test the signifance of this interaction, I fit the prediction data with a logistic mixed-effects regression model, using the lme4 package in R with the binomial link function. The model included random effects for subjects and sequences on the intercept, with fixed effects for Grammar (finite-state vs. context-free) and Prediction Length (one vs. three). The model confirmed that participants were more likely to guess the next word correctly when predicting a single word than when predicting three words, and that this effect was signifant ($\beta = 0.51$, $z = 3.746$, $p < 0.001$). There was also a signifcant interaction between Grammar and Prediction Length ($\beta = 0.41$, $z = 2.099$, $p < 0.05$), indicating that the main effect of Prediction Length was significantly greater among finite-state learners than among context-free learners. The effect of Grammar was not significant ($p = 0.38$).

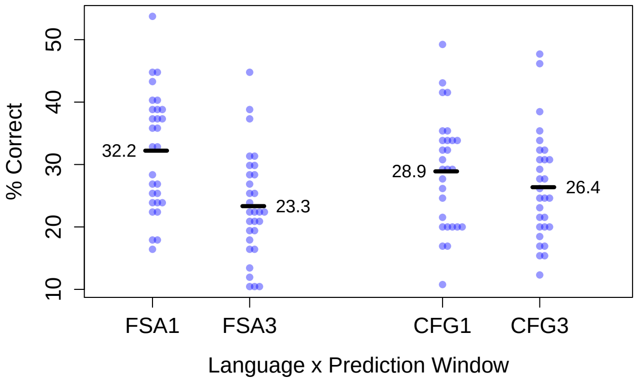

Figure 4: Accuracies in the final 40% of trials.

Adding fixed effects for Sequence Length and Prediction Position (measured as percentage of Length) revealed a significant effect of Position ($\beta = 1.25$, $z = 2.499$, $p \approx 0.01$). Participants were more likely to guess the answer correctly the later in the sequence they were asked to make their guess. These additional effect terms did not alter any of the results from the smaller model, though they did improve the overall model fit (ΔBIC $= 11.2$, $p < 0.05$).

3.2.3. Learning Over Time

The previous results paint a picture of the state of participants' respective language proficiencies towards the end of the task. But where the present design really shines is in its capacity to provide a window into the *dynamics* of language learning throughout the experiment.

The plot in Figure 5 displays moving averages for accuracies in the four conditions, collapsed across participants. Each point represents the collective accuracy of the participants within a condition on a particular trial. The lines trace out the average of the last 33 collective accuracies (thus the rolling window for accuracy accumulation consisted in $\frac{1}{5}$ of the total number of trials). Taken together, this gives an impression of the aggregate learning curves for the different conditions.

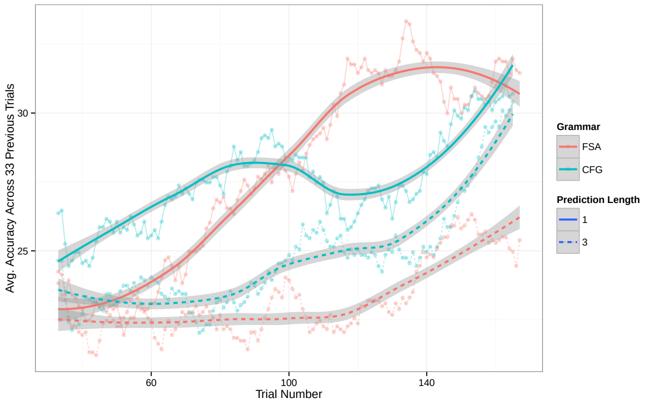

Figure 5: Learning curves for the four conditions. Each line represents the moving average of participant accuracies within a condition.

The most immediate property of these curves is that they are all somewhat different from one another. The predict-one FSA curve accelarates rapidly throughout the first two-thirds of the expeirment, but then plateaus (not altogether stably) for the rest of the experiment. In contrast, the predict-3 FSA curve is essentially flat for the duration of the initial period in which the FSA1 curve is accelerating upwards. That is, learning in this condition is all but absent for the first two-thirds of the experiment, but then increases gradually throughout. After drifting apart for the initial 100 trials, the two CFG curves both experience a sharp phase of acceleration late in the experiment, converging in the end toward the high accuracy of the FSA1 condition.

To statistically verify the difference in these trajectories, I fit logistic mixed-effects models to the entire data set, concentrating on interactions between Grammar, Prediction Length, and Time. To simplify the analysis, trials were divided into three stages of equal length. This division was primarily made to ensure sufficiently many data points to maintain statistical power while providing a dynamic component to the model, but can also be loosely justified by three apparent regions of continuous activity in Figure 5.

As before, the first model included random intercepts for subjects and sequences, as well as a fixed effects for Grammar, Prediction Length, and their interaction. In addition, it included a fixed effect for Stage (encoded as 1, 2, or 3, representing the first, second, and final thirds of the experiment, respectively). Across all four conditions, participants' accuracies increased over time, so it is unsurprising that the main effect of Stage is highly significant in this model ($\beta = 0.14$, $z = 6.669$, $p < 0.001$). Additionally, Prediction Length was a significant overall predictor of performance, with participants asked to predict a single word more likely to guess correctly than participants asked to predict the next three words ($\beta = -0.30$, $z = -3.239$, $p \approx 0.001$). Neither Grammar nor the interaction between Grammar and Prediction Length was significant, when data were collapsed across the three Stages.

To confirm the differential impact that Prediction Length played on the trajectory of learning for participants exposed to different Grammars, I added to the previous model a fixed three-way interaction effect for Grammar, Prediction Length, and Stage. That is, while there was no *overall* interaction between Grammar and Prediction Length, the analyses from the previous section showed that there was an interaction between Grammar and Prediction Length in the final 40% of trials. This shift is expected, since at the beginning of the experiment, participants in the four conditions should have performed fairly similarly, since they were all working with limited exposure to their respective languages. However, over the course of the experiment, we expect to see the influence of Prediction Length drive the two sets of Grammar learners in different directions. And indeed, this is what the model shows. The three-way interaction between

Grammar, Prediction Length, and Stage is significant ($\beta = 0.19$, $z = 2.284$, $p < 0.05$), as are the two-way interactions between Grammar and Stage ($\beta = -0.16$, $z = -2.822$, $p < 0.005$), and Prediction Length and Stage ($\beta = -0.17$, $z = -2.911$, $p < 0.005$).

3.3. Discussion

From the standpoint of my hypothesis, the most surprising result is that 3-predicting participants did worse than their 1-predicting counterparts, regardless of the language they made predictions about. I expected the length of the prediction window to be irrelevant for the finite-state learners, and to pull in precisely the opposite direction for those learning the context free language.

However, it is important to keep in mind that from an information-theoretic standpoint there was no difference between the FSA1 and FSA3 conditions. For instance, consider the example trial from Figure 3, 'sib daz jes tid | _____'. An FSA1 participant would be asked to provide a guess as to the next word in the sequence, while an FSA3 participant would be asked to provide a guess as to the next three words. But only the first word of the FSA3 participant's guess was scored, to equate the analysis of the two conditions. And in either condition, the entire rest of the sentence would be revealed after the participant made their guess(es). That means that despite having seen exactly the same set of prior sentences,[2] and making a prediction based on the same initial sentence fragment, the FSA1 participants were significantly more likely to answer correctly than the FSA3 participants.

This suggests that there was an unanticipated cost to the 3-prediction task. One potential source of the cost could have been the decreased levels of positive feedback. Recall that the storyteller only reacted positively if the participant exactly identified the relevant missing portion of the sentence. Even with perfect knowledge of either grammar, this required a bit of luck, as unfinished sentences could be both structurally and lexically ambiguous. And since within a categories, lexical ambiguities were independent of one another (for instance, though there is a syntactic dependency between the A category and the N category, there is no lexical dependency on the two A items and two N items), making it more unlikely that a participant would correctly identify the next three items than the next single item. This lack of positive feedback might have discouraged participants, decreasing engagement with the task. Or worse, it may have led them to reject good hypotheses about the *syntactic structure* of the language based on irrelevant lexical ambiguity more often than 1-predictors were led to do so.

Yet, the learning curves reveal that after a long period of stagnation — roughly half of the experiment — both FSA3 and CFG3 participants do show clear evidence of learning. And crucially, the late-stage improvements were not uniform between the two groups. In the second half of the experiment, the CFG3 learners nearly closed the gap created by the cost of the more difficult task. That is, in the final 20% of trials, the performance penalty associated with making longer predictions is effectively neutralized among the context-free learners, whose final performace is not significantly different from either the CFG1 learners ($t = 0.5427$, $p = 0.59$) or the FSA1 learners ($t = 0.8944$, $p = 0.37$). The FSA3 learners also improved in the last 20% of trials, but not at the same rate as the CFG3 learners [need growth rate analysis here].

4. General Discussion

One interpretation of this result is that while there was an initial, overall cost to the 3-prediction task, that cost was eventually offset by the benefit of the additional prediction-making for the participants attempting to learn the context free language. But participants attempting to learn the finite state language, for which there was no advantage to long-distance predictions (because there were no long-distance dependencies in the language), continued to suffer from the difficulty of the task. If this is indeed what is happening, then it provides indirect support for the hypothesis that long-distance prediction facilitates learning natural language-like sequential patterns — i.e. patterns with long-distance dependencies between implicit categories of vocabulary items — to a greater extent than

[2] Or at least, not significantly different prior sets of sentences. In actuality, the presentation of the stimuli were random, so averaged across conditions, the linguistic histories were effectively equivalent

it facilitates learning patterns with only first-order dependencies.

However, the data do not provide any evidence that deep predictions actually benefit language learning. On the contrary, for most of the experiment, those participants forced to make deep predictions performed significantly *worse* than those learning the same language from the same data making only immediate predictions. At best, we can conclude only that deep predictions eventually hinder context-free learning less than they hinder finite-state learning. That is, within the time frame of the experiment, context-free learners were able to overcome the cost of predicting three words into the future, while finite-state learners were not. As just pointed out, this is consistent with there being a benefit to long-distance prediction when mastering long-distance dependencies. But whatever benefit there is, it is inessential for achieving the prediction accuracy that the CFG3 participants achieved, given that the CFG1 participants achieved a similar, in fact, higher accuracy without making (explicit) long-distance predictions.

It is important to notice though that neither of the CFG conditions, nor the FSA3 condition, reached a period of stability in prediction accuracy. Participants were clearly still learning at the end of the 165 trials. This is most evident in the CFG conditions, where accuracies took a dramatic turn for the better nearly $\frac{2}{3}$ of the way through the experiment, and showed no signs of slowing down. It is possible that with more trials, the benefits of long-distance prediction would emerge, so that the CFG3 curve would overtake the CFG1 curve before they both plateaued. I leave the exploration of prediction length and its interaction with grammar under longer learning times to future investigation.

5. Conclusion

In this paper, I have been interested in the role that prediction and feedback play in language learning. Eye tracking and neuroimaging studies have revealed beyond a doubt that people constantly and incrementally anticipate what other people will say (Kamide 2008, Kutas et al. 2011). What's more, *errors* in prediction are strongly correlated with a variety of online processing measures, including reading times (Ehrlich and Rayner 1981, Levy 2008), syntactic priming effects (Jaeger and Snider 2013), and of course looking times (Tanenhaus et al. 1995, Altmann and Kamide 1999).

Indeed, several recent language learning and language adaptation models identify expectation-based revision as the primary mechanism by which learners develop and fine-tune their internal representation of the language they are learning (e.g. Chang et al. 2006, 2012, Pickering and Garrod 2013, Fine and Jaeger 2013). According to these theories, by making predictions about upcoming syntactic material, learners have the opportunity to formulate implicit hypotheses about the structure of their language, and to test those hypotheses in a never-ending observational experiment. Incorrect guesses lead the learner to adjust his or her underlying language model, either by redistributing probability mass over hypotheses that were compatible with the initial input, or by adjusting the hypothesis space itself to make room for the observed data.

These expectation-based processing and learning theories differ in how they represent natural language grammars, and therefore how they determine and redistribute probabilities in light of evidence. But despite these important differences, many of these models assume that individuals are one way or another calculating the likelihood of upcoming material by implicitly projecting all possible continuations of the current signal well into the future of the sentence. As reviewed in §2.2, there is even growing evidence that online processing costs are best explained by the aggregate uncertainty among all possible *complete sentences* left open by the unfolding fragment (Hale 2006, Frank 2013, Linzen and Jaeger 2014).

Despite the prominent position of anticipation and long-distance forward modeling in current theories of sentence processing, and the critical role of prediction error in emerging computational theories of language adaptation, there is at present effectively no direct experimental research into the question of *how* expectation formation affects language learning. As a first step in the direction of this important question, I hypothesized that long-distance prediction making would facilitate sequential learning, especially for complex patterns of the sort found in natural languages.

To test this, I conducted an experiment that invited participants to play the role of a cartoon character hearing a story in a language that it understood (but that the participant did not). From time to time, the participant's character would interrupt the storyteller, at which point the participant was asked to make a prediction about how the current sentence would continue. According to the hypothesis, participants being exposed to a context free language (with recursive, long-distance dependencies) should show greater or faster learning when asked to make predictions about a greater amount of upcoming sentence material. But participants exposed to a finite state language (with only local dependencies) should not show the same learning sensitivity to longer prediction windows.

I found that in the final 40% of trials, after a substantial amount of exposure, prediction, and feedback, there was indeed an interaction between grammar type and prediction length. In this phase of the data, the average CFG3 learner increased in accuracy at a faster rate than the average FSA3 learner. However, there was an overall, unexpected cost to predicting three items into the future, which dragged both the 3-predicting subgroups down below their 1-predicting counterparts. This means that there was no direct evidence for a language-learning benefit associated with lengthy predictions, even though the significant late-stage interaction between grammar type and prediction length suggests that the penalty for long predictions did not have quite the same effect on the CFG learners and teh FSA learners. In particular, CFG learners found it more helpful (or perhaps less harmful) to predict three words into the future than FSA learners. I then speculated that since three of the four experimental conditions continued to learn right up until the end of the experiment, it might be possible to see the hypothesized beneficial effect of long-distance predictions over a longer learning period, with enough trials to reach an accuracy ceiling.

6. References

Abrahamse, E. L., Jiménez, L., Verwey, W. B., Clegg, B. A., 2010. Representing serial action and perception. Psychonomic bulletin & review 17 (5), 603–623.

Alexandre, J. D., 2010. Modeling implicit and explicit processes in recursive sequence structure learning. In: The 32nd annual meeting of the Cognitive Science Society (CogSci10). Austin, TX: Cognitive Science Society. pp. 1381–1386.
URL http://csjarchive.cogsci.rpi.edu/proceedings/2010/papers/0379/paper0379.pdf

Altmann, G., Kamide, Y., 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition 73 (3), 247–264.
URL http://www.sciencedirect.com/science/article/pii/S0010027799000591

Archer, E., 1960. Re-evaluation of the Meaningfulness of All Possible CVC Trigrams. Psychological monographs: general and applied. American Psychological Association.
URL http://books.google.com/books?id=QP3COwAACAAJ

Aslin, R. N., Saffran, J. R., Newport, E. L., 1998. Computation of conditional probability statistics by 8-month-old infants. Psychological science 9 (4), 321–324.
URL http://pss.sagepub.com/content/9/4/321.short

Blache, P., Rauzy, S., 2011. Predicting linguistic difficulty by means of a morpho-syntactic probabilistic model. In: PACLIC. Vol. 25. p. 160–167.
URL http://www.aclweb.org/anthology-new/Y/Y11/Y11-1017.pdf

Chang, F., Dell, G. S., Bock, K., 2006. Becoming syntactic. Psychological review 113 (2), 234.
URL http://psycnet.apa.org/journals/rev/113/2/234/

Chang, F., Janciauskas, M., Fitz, H., 2012. Language adaptation and learning: Getting explicit about implicit learning. Language and Linguistics Compass 6 (5), 259–278.
URL http://onlinelibrary.wiley.com/doi/10.1002/lnc3.337/full

Chomsky, N., 1956. Three models for the description of language. Information Theory, IRE Transactions on 2 (3), 113–124.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1056813

Christiansen, M. H., Louise Kelly, M., Shillcock, R. C., Greenfield, K., 2010. Impaired artificial grammar learning in agrammatism. Cognition 116 (3), 382–393.

Clark, A., 2013. Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences 36 (03), 181–204.

Clegg, B. A., DiGirolamo, G. J., Keele, S. W., 1998. Sequence learning. Trends in cognitive sciences 2 (8), 275–281.

Conway, C. M., Bauernschmidt, A., Huang, S. S., Pisoni, D. B., Mar. 2010. Implicit statistical learning in language processing: Word predictability is the key. Cognition 114 (3), 356–371.
URL http://ezproxy.library.nyu.edu:2060/science/article/pii/S0010027709002613#

Creel, S. C., Newport, E. L., Aslin, R. N., 2004. Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition 30 (5), 1119–1130.
URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.30.5.1119

De Vries, M. H., Barth, A. C., Maiworm, S., Knecht, S., Zwitserlood, P., Flöel, A., 2010. Electrical stimulation of broca's area enhances implicit learning of an artificial grammar.

Journal of Cognitive Neuroscience 22 (11), 2427–2436.

de Vries, M. H., Monaghan, P., Knecht, S., Zwitserlood, P., May 2008. Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. Cognition 107 (2), 763–774.
URL http://www.sciencedirect.com/science/article/pii/S0010027707002533

de Vries, M. H., Petersson, K. M., Geukes, S., Zwitserlood, P., Christiansen, M. H., Jul. 2012. Processing multiple non-adjacent dependencies: evidence from sequence learning. Philosophical Transactions of the Royal Society B: Biological Sciences 367 (1598), 2065–2076, PMID: 22688641.
URL http://rstb.royalsocietypublishing.org/content/367/1598/2065

DeLong, K. A., Urbach, Thomas P. andnd Kutas, M., 2005. Probabilistic word probabilisticre-activation during language comprehension inferred from electrical brain activity. Nature neuroscience 8 (8), 1117–1121.
URL http://www.nature.com/neuro/journal/v8/n8/abs/nn1504.html

Ehrlich, S. F., Rayner, K., 1981. Contextual effects on word perception and eye movements during reading. Journal of verbal learning and verbal behavior 20 (6), 641–655.
URL http://www.sciencedirect.com/science/article/pii/S0022537181902206

Elman, J. L., 1990. Finding structure in time. Cognitive science 14 (2), 179–211.
URL http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1402_1/abstract

Federmeier, K. D., Jul. 2007. Thinking ahead: The role and roots of prediction in language comprehension. Psychophysiology 44 (4), 491–505.
URL http://ezproxy.library.nyu.edu:2153/doi/10.1111/j.1469-8986.2007.00531.x/abstract?deniedAccessCustomisedMessage=&userIsAuthenticated=false

Fine, A. B., Jaeger, T. F., 2013. Evidence for implicit learning in syntactic comprehension. Cognitive Science 37 (3), 578–591.
URL http://onlinelibrary.wiley.com/doi/10.1111/cogs.12022/full

Fitch, W. T., Hauser, M. D., 2004. Computational constraints on syntactic processing in a nonhuman primate. Science 303 (5656), 377–380.
URL http://www.sciencemag.org/content/303/5656/377.short

Frank, S. L., 2010. Uncertainty reduction as a measure of cognitive processing effort. In: Proceedings of the 2010 workshop on cognitive modeling and computational linguistics. Association for Computationalional Linguistics, p. 81–89.
URL http://dl.acm.org/citation.cfm?id=1870075

Frank, S. L., Jul. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. Topics in Cognitive Science 5 (3), 475–494.

Gebhart, A. L., Newport, E. L., Aslin, R. N., 2009. Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. Psychonomic bulletin & review 16 (3), 486–490.
URL http://link.springer.com/article/10.3758/PBR.16.3.486

Gentner, T. Q., 2007. Mechanisms of temporal auditory pattern recognition in songbirds. Language learning and development 3 (2), 157–178.
URL http://www.tandfonline.com/doi/full/10.1080/15475440701225477

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

Gomez, R. L., Gerken, L., Schvaneveldt, R. W., Mar. 2000. The basis of transfer in artificial grammar learning. Memory & Cognition 28 (2), 253–263.
URL http://link.springer.com/article/10.3758/BF03213804

Gureckis, T. M., Love, B. C., 2010. Direct associations or internal transformations? exploring the mechanisms underlying sequential learning behavior. Cognitive Science 34 (1), 10–50.
URL http://doi.wiley.com/10.1111/j.1551-6709.2009.01076.x

Gómez, R., Maye, J., 2005. The developmental trajectory of nonadjacent dependency learning. Infancy 7 (2), 183–206.
URL http://onlinelibrary.wiley.com/doi/10.1207/s15327078in0702_4/abstract

Hale, J., 2001. A probabilistic earley parser as a psycholinguistic model. In: Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics, pp. 1–8.

Hale, J., 2006. Uncertainty about the rest of the sentence. Cognitive Science 30 (4), 643–672.
URL http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog0000_64/abstract

Hochmann, J.-R., Azadpour, M., Mehler, J., 2008. Do humans really learn AnBn artificial grammars from exemplars? Cognitive Science 32 (6), 1021–1036.
URL http://onlinelibrary.wiley.com/doi/10.1080/03640210801897849/abstract

Jaeger, T. F., Snider, N. E., 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. Cognition 127 (1), 57–83.
URL http://www.sciencedirect.com/science/article/pii/S0010027712002636

Jurafsky, D., 1996. A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science volume20 (2), 137–194.
URL http://onlinelibrary.wiley.com/doi/10.1207/s15516709cog2002_1/abstract

Kamide, Y., Jul. 2008. Anticipatory processes in sentence processing. Language and Linguistics Compass 2 (4), 647–670.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2008.00072.x/abstract

Kutas, M., DeLong, K. A., Smith, N. J., 2011. A look around at what lies ahead: Prediction and predictability in language processing. Predictions in the brain: Using our past to generate a future, 190–207.

Lai, J., Poletiek, F. H., Feb. 2011. The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. Cognition 118 (2), 265–273.
URL http://ezproxy.library.nyu.edu:2060/science/article/pii/S0010027710002829

Lerdahl, F., Jackendoff, R. S., 1983. A generative theory of tonal music. MIT press.

Levy, R., Mar. 2008. Expectation-based syntactic comprehension. Cognition 106 (3), 1126–1177.
URL http://www.sciencedirect.com/science/article/pii/S0010027707001436

Levy, R., 2013. Memory and surprisal in human sentence comprehension. Sentence Processing, 78.

Linzen, T., Jaeger, T. F., 2014. Investigating the role of entropy in sentence processing. ACL 2014, 10.
URL http://acl2014.org/acl2014/W14-20/W14-20-2014.pdf#page=20

Mani, N., Huettig, F., 2012. Prediction during language processing is a piece of cake—But only for skilled producers. Journal of Experimental Psychology: Human Perception and Performance 38 (4), 843–847.

Marcus, G. F., Vijayan, S., Rao, S. B., Vishton, P. M., 1999. Rule learning by seven-month-old infants. Science 283 (5398), 77–80.
URL http://www.sciencemag.org/content/283/5398/77.short

Maye, J., Werker, J. F., Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82 (3), B101–B111.
URL http://www.sciencedirect.com/science/article/pii/S0010027701001573

Misyak, J. B., Christiansen, M. H., Bruce Tomblin, J., 2010. Sequential expectations: The role of prediction-based learning in language. Topics in Cognitive Science 2 (1), 138–153.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.1756-8765.2009.01072.x/abstract

Nissen, M. J., Bullemer, P., 1987. Attentional requirements of learning: Evidence from performance measures. Cognitive psychology 19 (1), 1–32.

Opitz, B., Friederici, A. D., Sep. 2004. Brain correlates of language learning: The neuronal dissociation of rule-based versus similarity-based learning. The Journal of Neuroscience 24 (39), 8436–8440, PMID: 15456816.
URL http://www.jneurosci.org/content/24/39/8436

Opitz, B., Friederici, A. D., 2007. Neural basis of processing sequential and hierarchical syntactic structures. Human Brain Mapping 28 (7), 585–592.
URL http://onlinelibrary.wiley.com/doi/10.1002/hbm.20287/abstract

Perruchet, P., Rey, A., 2005. Does the mastery of center-embedded linguistic structures distinguish humans from non-human primates? Psychonomic Bulletin & Review 12 (2), 307–313.

Petersson, K. M., Hagoort, P., Jul. 2012. The neurobiology of syntax: beyond string sets. Philosophical Transactions of the Royal Society B: Biological Sciences 367 (1598), 1971–1983, PMID: 22688633.
URL http://rstb.royalsocietypublishing.org/content/367/1598/1971

Pickering, M. J., Garrod, S., Aug. 2013. An integrated theory of language production and comprehension. Behavioral and Brain Sciences 36 (04), 329–347.
URL http://ezproxy.library.nyu.edu:2102/action/displayFulltext?type=1&fid=8958099&jid=BBS&volumeId=36&issueId=04&aid=8958097&bodyId=&membershipNumber=&societyETOCSession=

Pothos, E. M., 2007. Theories of artificial grammar learning. Psychological Bulletin 133 (2), 227–244.

Reber, A. S., Dec. 1967. Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior 6 (6), 855–863.
URL http://www.sciencedirect.com/science/article/pii/S002253716780149X

Rescorla, R. A., Wagner, A. R., 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: Current research and theory 2, 64–99.

Roark, B., Bachrach, A., Cardenas, C., Pallier, C., 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. Association for Computational Linguistics, p. 324–333.
URL http://dl.acm.org/citation.cfm?id=1699553

Rohrmeier, M., 2011. Towards a generative syntax of tonal harmony. Journal of Mathematics and Music 5 (1), 35–53.
URL http://www.tandfonline.com/doi/abs/10.1080/17459737.2011.573676

Rohrmeier, M., Fu, Q., Dienes, Z., Oct. 2012. Implicit learning of recursive context-free grammars. PLoS ONE 7 (10), e45885.
URL http://dx.doi.org/10.1371/journal.pone.0045885

Romberg, A. R., Saffran, J. R., 2012. Expectancy learning from probabilistic input by infants. Frontiers in psychology 3, 1–16.

Saffran, J., Jul. 2002. Constraints on statistical language learning. Journal of Memory and Language 47 (1), 172–196.
URL http://ezproxy.library.nyu.edu:2060/science/article/pii/S0749596X01928396

Saffran, J. R., Aslin, R. N., Newport, E. L., 1996a. Statistical learning by 8-month-old infants. Science 274 (5294), 1926–1928.
URL http://www.sciencemag.org/content/274/5294/1926.short

Saffran, J. R., Newport, E. L., Aslin, R. N., 1996b. Word segmentation: The role of distributional cues. Journal of

memory and language 35 (4), 606–621.
URL http://www.sciencedirect.com/science/article/pii/S0749596X96900327

Santelmann, L. M., Jusczyk, P. W., 1998. Sensitivity to discontinuous dependencies in languageanguage learners: Evidence for limitations in processing space. Cognition 69 (2), 105–134.
URL http://www.sciencedirect.com/science/article/pii/S0010027798000602

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., Sedivy, J. C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268 (5217), 1632–1634.
URL http://www.sciencemag.orgg/content/268/5217/1632.short

Thompson, S. P., Newport, E. L., 2007. Statistical learning of syntax: The role of transitional probability. Language Learning and Development 3 (1), 1–42.
URL http://www.tandfonline.com/doi/full/10.1080/15475440709336999

Trueswell, J. C., Medina, T. N., Hafri, A., Gleitman, L. R., Feb. 2013. Propose but verify: Fast mapping meets cross-situational word learning. Cognitive Psychology 66 (1), 126–156.
URL http://ezproxy.library.nyu.edu:2060/science/article/pii/S0010028512000795

Trueswell, J. C., Tanenhaus, M. K., Garnsey, S. M., 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of memory and language 33 (3), 285–318.
URL http://www.httpsciencedirect.com/science/article/pii/S0749596X8471014X

Uddén, J., Ingvar, M., Hagoort, P., Petersson, K. M., 2012. Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: Investigating the push-down stack model. Cognitive Science 36 (6), 1078–1101.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2012.01235.x/abstract

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. Journal of Experimental Psychology: Learningg, Memory, and Cognition 31 (3), 443.
URL howttp://psycnet.apa.org/journals/xlm/31/3/443/

Vitevitch, M. S., Luce, P. A., 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. Journal of Memory and Language 40 (3), 374–408.
URL http://www.sciencedirect.com/science/article/pii/S0749596X98926183

Wu, S., Bachrach, A., Cardenas, C., Schuler, W., 2010. Complexity metrics in an incremental right-corner parser. In: Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, p. 1189–1198.
URL http://dl.acm.org/citation.cfm?id=id1858802