LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

A Role for Chunk Formation in Statistical Learning of Second Language Syntax

Phillip Hamrick

Kent State University

Humans are remarkably sensitive to the statistical structure of language. However, different mechanisms have been proposed to account for such statistical sensitivities. The present study compared adult learning of syntax and the ability of two models of statistical learning to simulate human performance: Simple Recurrent Networks, which learn by predictive computation, and PARSER, which learns chunks as a byproduct of general principles of associative learning and memory. In the first stage, a semiartificial language paradigm was used to gather human data. In the second stage, a simulation paradigm was then used to compare the patterns of performance of the SRN and PARSER. After the human adults and the computational models were trained on sentences from the semiartificial language with probabilistic syntax, their learning outcomes were compared. Neither model was able to fully reproduce the human data, which may indicate less robust statistical learning effects in adults; however, PARSER was able to simulate more of the adult learning data than the SRN, suggesting a possible role for chunk formation in early phases of adult learning of second language syntax.

Keywords statistical learning; second language acquisition; syntax; chunking; PARSER; Simple Recurrent Network

Introduction

There is increasing evidence that humans are sensitive to the statistical properties inherent in language (for overviews, see Rebuschat & Williams, 2012; Romberg & Saffran, 2010). Evidence for such sensitivity in infants and adults comes from speech segmentation (e.g., Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996), word learning (e.g., Yu & Smith, 2007),

Thanks to Pierre Perruchet, Rebecca Sachs, John Williams, and the anonymous reviewers for their very insightful critiques and comments. Many thanks also to Patrick Rebuschat, Cristina Sanz, Alison Mackey, Michael Long, and Nick Ellis for their input at various stages in the development of this study. The data, findings, and conclusions are solely the responsibility of the author.

Correspondence concerning this article should be addressed to Phillip Hamrick, Language and Cognition Research Laboratory, Department of English, 113 Satterfield Hall, Kent State University, Kent, OH, USA 44242. E-mail: phamric1@kent.edu.

Language Learning 64:2, June 2014, pp. 247–278

DOI: 10.1111/lang.12049

artificial grammar learning (e.g., Gomez & Gerken, 1999), and phrase structure and syntax learning (e.g., Saffran, 2001; Thompson & Newport, 2007). The robustness of these findings has led many to posit that humans possess powerful statistical learning mechanisms (e.g., Aslin et al., 1998; Gopnik, Wellman, Gelman, & Meltzoff, 2010; Saffran, 2003) capable of extracting a variety of types of knowledge.

However, despite considerable evidence demonstrating the robustness of statistical learning, there is little agreement regarding the nature of the cognitive mechanisms that underlie statistical learning. One common conceptualization of statistical learning mechanisms is as statistical computation (e.g., Aslin et al., 1998). What statistical computation entails is a matter of debate. Some researchers posit that human cognition is endowed with the ability to make powerful, complex statistical inferences unconsciously (e.g., Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). However, it seems unlikely that such implicit inferences are the same as those that a statistician would make consciously. Therefore, although computational models operating under this assumption have been able to simulate a variety of human behaviors, they have been criticized for their lack of constraint and psychological plausibility (e.g., Altmann, 2010; Perruchet & Peereman, 2004, p. 98). More psychologically plausible approaches suggest that unconscious statistical computations can be thought of as tracking predictive dependencies or as learning probabilistic cue-outcome relationships (i.e., contingency learning). Connectionist networks provide one way of simulating these sorts of computations (e.g., Misyak, Christiansen, & Tomblin, 2010; Shanks, 1995; Williams, 2009).

However, this is not the only psychologically plausible conceptualization of statistical learning. An alternative explanation is that sensitivity to frequencies and probabilities in language are due not to a predictive statistical learning mechanism, but rather to chunk formation. This notion is far from new. For example, researchers in the implicit learning paradigm have long regarded chunk formation as one way in which learners might develop sensitivities to statistics (e.g., Knowlton & Squire, 1996; Perruchet & Pacteau, 1990; Reber & Lewis, 1977; Redington & Chater, 1996; Servan-Schreiber & Anderson, 1990). Although these approaches have focused primarily on how chunk formation exploits surface frequency information, more recent models, like PARSER (Perruchet & Vinter, 1998), have considered how chunking may result in sensitivity to more than just frequency. For example, in their PARSER model, Perruchet and Vinter (1998) argue that phenomenal awareness is the starting point of learning (cf. Schmidt, 1990). The attended content of phenomenal awareness forms a chunk. This chunk forms a memory trace and this is subsequently strengthened (i.e., through repeated exposure) or weakened (i.e., through decay and interference). This process leads to the emergence of chunk knowledge that is increasingly well matched to the statistical regularities in the language (Perruchet & Peereman, 2004). This process is described in more detail later in this article.

The aim of the present study was to investigate which of these statistical learning mechanisms (statistical computation or chunk formation) better account for adult performance in a relatively new domain of statistical learning research: incidental learning of second language (L2) syntax. As it stands, there is currently no consensus regarding the mechanisms of statistical learning in L2 development (or in other domains, for that matter). Some research has considered chunk formation to be central (e.g., Ellis, 1996, 2003; Robinson, 2005) while other research has invoked statistical computation in the connectionist sense (e.g., Ellis & Schmidt, 1997; Williams, 2010; Williams & Kuribara, 2008). However, to the best of my knowledge, no L2 study has focused on assessing these different mechanisms by comparing the competing predictions of different computational models of statistical learning with actual learning in adults. To that end, the present study compared the predictions of two classes of statistical learning mechanisms with adult learning of syntax in a semiartificial language. Following previous work on statistical learning and L2 syntax by Williams (2010; Williams & Kuribara, 2008), the first class of mechanism is one that computes predictive statistics (Simple Recurrent Network [SRN]; e.g., Elman, 1990; Williams, 2010). The second class of mechanism is one that learns via chunk formation (PARSER; Perruchet & Vinter, 1998). Both of these models have been used to account for various psycholinguistic phenomena, but PARSER has not previously been used in L2 research. Instead, PARSER has been used to account for child and adult acquisition patterns in word segmentation (e.g., Giroux & Rey, 2009; Perruchet & Vinter, 1998) and syllable processing (Perruchet & Peereman, 2004). On the other hand, the SRN model has been one of the most widely used computational models of statistical learning and has been remarkably successful in a variety of linguistic domains, including syntax (e.g., Chang, Dell, & Bock, 2006; Christiansen & Chater, 2001; Williams, 2010; Williams & Kuribara, 2008) and spoken-word recognition (e.g., Gaskell & Marslen-Wilson, 2001).

Two Computational Models of Statistical Learning

In this section I briefly review the details of the SRN and PARSER. For each, I discuss its architecture and mechanisms, followed by an illustration, and concluding with a brief discussion of previous findings using the models.

SRN

Perhaps the most widely used computational model of statistical learning has been the SRN, which was proposed by Elman (1990). As with all connectionist networks, learning in the SRN takes place via the progressive tuning of connections between the layers of units in the network, and input representations are mapped onto output representations through hidden units. However, the SRN has an extra "copy" or "context" layer, which provides the network with a feedback loop of information. This feature gives the network a sort of short-term memory for preceding material, allowing it to associate previous input with current input. Together, these architectural qualities give the SRN the ability to learn by predictive computation: its ability to learn by predicting the next element in a sequence. In order to train the SRN, it is given a sequential input one item at a time. At each time step, the SRN makes a prediction about what comes next in a sequence. When the next item is input, the SRN compares that input with its previous prediction. If its prediction is correct, then the SRN adjusts its connection weights to increase the likelihood that it makes the same correct prediction in the future. If its prediction is incorrect, however, then the SRN adjusts its connection weights to increase the likelihood that it does not make the same inaccurate prediction in the future. It is this feature that makes the SRN sensitive to forward transitional probabilities, that is, the likelihood that one item follows another in a sequence.

To illustrate, let us apply the SRN to a simplified, constructed example: an unsegmented stream of letters composed of the bigrams AB, CD, EF, and GH. Assume that the stream consists of the following sequence: *ABCDEFGHCDGHEFABEFCDABGH*. The transitional probabilities within bigrams are all equal to one, for example, A is followed by B 100% of the time, C is followed by D 100% of the time, and so on. However, the transitional probabilities *between* bigrams are lower. For instance, D is followed by E 33% of the time, G 33% of the time, and A 33% of the time. For the sake of illustration, let us imagine that when the SRN is first exposed to the letter A, it predicts that C comes next. However, upon inputting the second letter, B, the network must correct for its inaccurate prediction. It does so by changing its connection weights such that the next time it encounters A, it will be less likely to predict that C comes next. By repeating this process over a large enough corpus, the SRN will learn to approximate the low transitional probabilities between bigrams and the high transitional probabilities within bigrams. In the context of the present example, this means that, if the network has learned, when it is presented with A it will correctly predict that B comes next. In essence, it will have learned the AB bigram. This approach is often referred to as bracketing because it assumes that humans insert boundaries between items that have a low transitional probability, with sequences containing higher transitional probabilities assumed to reflect whole units. In other words, chunked units are inferred only after statistics have been computed.

How does such a statistical learning mechanism compare with human learning in our present area of interest, L2 syntax? Williams and Kuribara (2008) and Williams (2010) addressed precisely this question in the context of adult learning of L2 syntax using a semiartificial language paradigm. In Williams and Kuribara's (2008) study, participants read sentences in Japlish, a semiartificial language consisting of English words placed into Japanese syntactic structures and affixed with Japanese morphemes (e.g., *John-ga pizza-o ate;* John-NOM pizza-ACC ate; John ate pizza). Participants read Japlish sentences under the guise of a plausibility judgment task and were then given a surprise grammaticality judgment task (GJT). Williams and Kuribara then showed that an SRN trained on the syntactic category sequences from Japlish was able to closely match human behavior on the GJT, accounting for between 69% and 91% of the adult learner data.

Extending this work further, Williams (2010) conducted an experiment using nonsense syllable classes (e.g., *si/se/sa/so* and *pi/pe/pa/po*) as analogs for Japlish syntactic categories (e.g., *Horse-ni farmer-ga hay-o gave* became *to-ni so-ga pa-o ku*). Thus, any nonsense syllable beginning with '*s*' corresponded to the subject of the sentence and so on. Again, participants were given a GJT to assess learning. Then, Williams again simulated human learning behavior by training a SRN, this time on the nonsense syllable class (i.e., syntactic category analog) sequences from the training phase of the second incidental learning experiment. The results showed that the SRN was able to account for approximately 96% of the variance in the human data in the experiment with the meaningless nonsense syllables, but only 40% and 66% of the data for participants trained on actual Japlish (instead of the analog). The reduction in fit was, presumably, due to the influence of other linguistic factors.

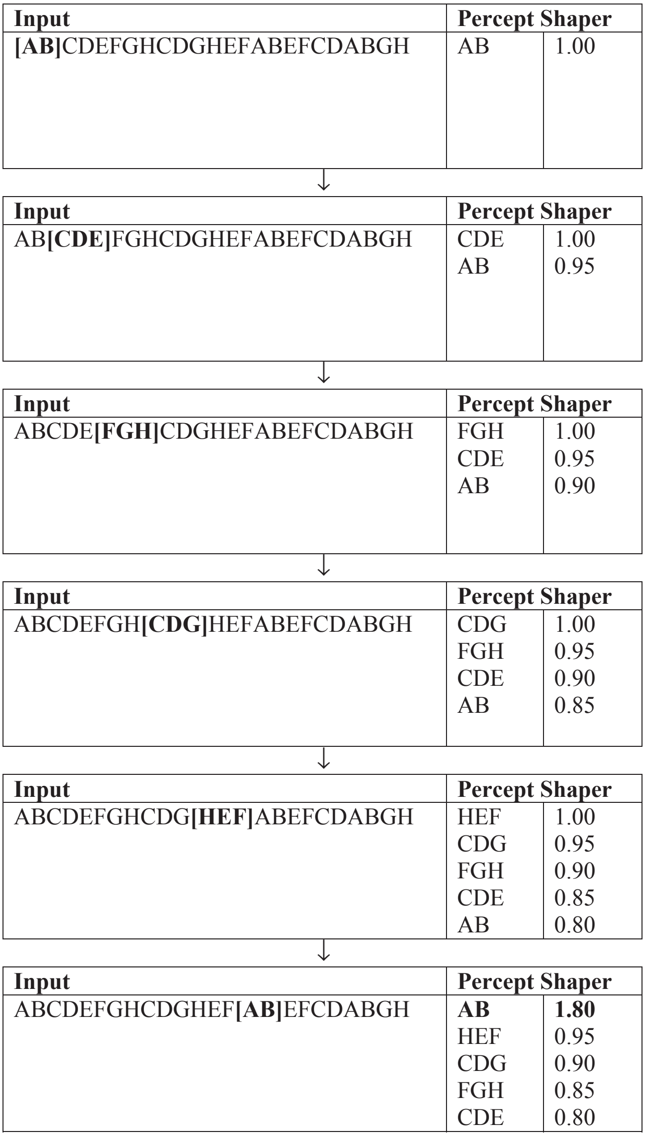
These results show that, in principle, the SRN can mimic human learning of syntactic patterns if it is trained on syntactic category sequences. A plausible interpretation of this finding is that humans possess mechanisms that are functionally comparable to the predictive mechanisms of the SRN, and such proposals exist (e.g., Altmann & Mirkovič, 2009). However, it is important to keep in mind that these studies comparing adult learning of L2 syntax and SRN performance were not aimed at elucidating the mechanisms of statistical learning, but were, rather, aimed at demonstrating that statistical learning mechanisms suffice to explain human learning in the behavioral experiments. Consequently, there were no comparisons between multiple models of statistical learning. This leaves open the possibility that other statistical learning mechanisms may also be able to explain the human data. That is, functional equivalence to the performance of the SRN may be obtained through a number of different mechanisms. There is some evidence to this effect. Perruchet and Peereman (2004) compared word-likeness ratings as a function of statistical information in French rimes (VC). They then compared human ratings with those predicted by the SRN and the chunk formation model PARSER. PARSER significantly correlated with the SRN ($r = .69$) and better predicted human ratings than the SRN. Thus, there is good reason to ask whether a different model like PARSER could account for adult learning of L2 syntax in the semiartificial paradigm described above without recourse to the predictive computations found in the SRN.

PARSER

Perruchet and Vinter (1998, 2002) presented PARSER as a new computational model to account for the word segmentation phenomena in the seminal studies conducted by Saffran and colleagues (e.g., Aslin et al., 1998; Saffran et al., 1996). PARSER's attractiveness is in its parsimony. It has been used to simulate findings in the statistical learning literature on word segmentation without the addition of specific mechanisms that perform statistical computations (e.g., Giroux & Rey, 2009; Perruchet & Tillman, 2010; Perruchet & Vinter, 1998). Instead, PARSER learns by a process of chunk formation that is grounded in well-established principles of associative learning and memory. First, chunks are formed initially on a random basis by the limited capacity of attention.[1] The content of PARSER's attentional focus forms a chunk that enters into its memory system (known as the percept shaper). The size of the chunk depends on the varying size of the attentional focus, which can be manipulated by the researcher for a priori theoretical reasons. Chunks are then gradually strengthened or weakened with further experience through general principles of associative memory decay and interference. Like the attentional focus parameter, the rate of memory decay and interference can be manipulated by the researcher for a priori theoretical reasons.

If the SRN represents a bracketing approach, then PARSER represents a clustering approach (Giroux & Rey, 2009). Processing primitives that occur within the window of attention are repeatedly clustered together forming increasingly complex chunks. To illustrate, let us also apply PARSER to the unsegmented stream of letters from the sequence above, again composed of the bigrams AB, CD, EF, and GH (see Figure 1 for an illustration). Initially, because PARSER does not know the structure of the input, it will extract random chunks from a series of attentional processing episodes. Chances are that some of these chunks will be legal bigrams, while others will not be. Consider a very simplified example. With no cues to segmentation, PARSER might extract the units AB, CDE, and FGH from the first part of the above artificial grammar string *ABCDEFGH*. In this case, PARSER has created units in memory for one correct bigram (AB) and two illegal trigrams (CDE and FGH). These units in memory now guide the processing of the next section of the above string: *CDGHEFAB*. Assume for the sake of simplicity that PARSER experiences this string as three more discrete chunks: CDG, HEF, and AB. In the processing of this second string, neither of the original illegal trigrams CDE or FGH has been repeated in the input, so their memory traces decay. Moreover, the fact that these trigrams overlap with the recently processed trigrams CDE and FGH means that the original illegal trigrams in memory are further forgotten due to interference. Finally, the legal bigram AB that was chunked in the first pass again matches the bigram AB in the second pass, leading to the strengthening of the AB bigram in PARSER's memory. As this process repeats on a large corpus, PARSER is able to converge on the correct units of the input, which, as noted above, have statistical structure.

Unlike the SRN, PARSER has not been used in second language acquisition (SLA) research. Rather, PARSER was designed to account for word segmentation phenomena in child language acquisition research on statistical learning (e.g., Saffran et al., 1996). It follows, then, that applying PARSER to syntax acquisition assumes a functional parallel between the mechanisms responsible for word segmentation and the mechanisms of syntactic development. For word segmentation, PARSER segments language into disjunctive parts composed of primitives (e.g., syllables or phonemes). Attentional processes and associative learning principles lead to the gradual emergence of clusters of primitives in memory (i.e., words and morphemes). In the present application to syntax, adults are assumed to have abstract knowledge of syntactic categories (or some comparable structures) as syntactic primitives, and PARSER operates on these abstract primitives. This is both a theoretical assumption and a methodological detail that maintains comparability between the present study with the work on L2 syntax using the SRN in Williams and Kuribara (2008) and Williams (2010). PARSER segments the input into disjunctive chunks of syntactic



Input	Percept Shaper	
[AB]CDEFGHCDGHEFABEFCDABGH	AB	1.00

↓

Input	Percept Shaper	
AB**[CDE]**FGHCDGHEFABEFCDABGH	CDE AB	1.00 0.95

↓

Input	Percept Shaper	
ABCDE**[FGH]**CDGHEFABEFCDABGH	FGH CDE AB	1.00 0.95 0.90

↓

Input	Percept Shaper	
ABCDEFGH**[CDG]**HEFABEFCDABGH	CDG FGH CDE AB	1.00 0.95 0.90 0.85

↓

Input	Percept Shaper	
ABCDEFGHCDG**[HEF]**ABEFCDABGH	HEF CDG FGH CDE AB	1.00 0.95 0.90 0.85 0.80

↓

Input	Percept Shaper	
ABCDEFGHCDGHEF**[AB]**EFCDABGH	**AB** HEF CDG FGH CDE	**1.80** 0.95 0.90 0.85 0.80

Figure 1 Schematic illustration of chunk formation in PARSER. Given the perceptual primitives A, B, C, D, E, F, and G, PARSER randomly forms chunks based on attentional windows of 2–3 letters in size. The chunks are strengthened or decayed with more input. The example uses PARSER's default attention window and memory decay settings.

primitives (e.g., NP-VP or PP-VP). The end result is an inventory of increasingly complex syntactic chunks. It is worth noting here that, although PARSER was not originally intended to account for syntactic development phenomena, there is no reason in principle it cannot apply to more abstract knowledge structures such as syntactic categories (Perruchet, 2005). Moreover, the iterative formation of increasingly large and complex syntactic chunks at least tacitly aligns the present use of PARSER with theoretical frameworks that treat syntactic development as at least in part the product of learning abstract syntactic chunks, including construction grammar (e.g., Goldberg, 2006; Tomasello, 2003), simpler syntax (e.g., Culicover & Jackendoff, 2006), the memory-unification framework (Hagoort, 2005), and head-driven phrase structure grammar (Pollard & Sag, 1994).

How do the SRN and PARSER compare in their ability to account for human learning phenomena? The answer is not clear. The SRN has been much more widely used than PARSER in simulating human learning and has been very successful in accounting for a variety of linguistic phenomena. On the other hand, PARSER has been shown to better and more parsimoniously account for some human linguistic data (e.g., Giroux & Rey, 2009; Perruchet & Peereman, 2004). For example, Giroux and Rey (2009) trained a SRN and PARSER on a word segmentation task (like that of the studies in Saffran et al., 1996) and compared the competing predictions of these models with human performance on the same task. When trained on the word segmentation task, the SRN predicts that, over time, word segmentation should lead to increased unit status for both lexical and sublexical units. On the same task, PARSER predicts that only words would be strengthened over time, with sublexical units being increasingly less weighted in memory than lexical units. The human data fell within the predictions of PARSER. With increased training, full words were learned better than sublexical units (i.e., part-words), consistent with PARSER. This finding is broadly consistent with other studies showing that other chunk-based models outperform other statistical sequence learning models in other domains, such as visual pattern learning (e.g., Orbán, Fiser, Aslin, & Lengyel, 2008) and artificial grammar learning (e.g., Boucher & Dienes, 2003). Importantly, this suggests that advantages for chunk-based models over sequence learning models are not limited to the idiosyncrasies of a single model (i.e., PARSER) in a single domain (i.e., speech segmentation), but instead may be more general property of chunk formation itself as a mechanism of human learning.

The Present Study

The study reported in this article extends these previous investigations comparing the performance of different computational models of statistical learning, focusing specifically on the SRN and PARSER. In the first stage, a semiartificial language paradigm similar to that of Williams (2010; Williams & Kuribara, 2008) was used to gather human data on incidental learning of L2 syntax. In the second stage, a simulation paradigm was used to compare the behavioral data with the pattern of performance in the SRN and PARSER in order to see which computational model, if any, was better able to capture the human pattern of performance.

The study contained three innovations in the semiartificial language learning paradigm. First, it used a trained control group in order to avoid problematic assumptions involved in comparing experimental groups with untrained controls or chance baselines (e.g., Hamrick, 2012, 2013; Rebuschat, Hamrick, Sachs, Riestenberg, & Ziegler, 2013; Perruchet & Reber, 2003; see Dienes & Altmann, 2003, for a counterargument). The second methodological innovation concerned the development of the semiartificial language stimuli. In previous studies using semiartificial languages, the statistical structure of the stimuli was not controlled. So, even if statistical learning took place, it was unclear what types of statistics were informative for learners. Therefore, the present study employed a semiartificial language whose syntactic structure was constructed to have only two transitional probabilities between syntactic categories (67% and 33%). The third innovation was that results of the human experiment were compared with two computational models of statistical learning, rather than one: the SRN, which learns via predictive computation, and PARSER, which learns via chunk formation.

Stage 1: Behavioral Evidence

In order to address the above goals, first a semiartificial language learning experiment was conducted on human adults.[2] The sections below report on the method and findings obtained in this first stage of the study.

Method

Participants

Thirty volunteer undergraduate native speakers of English (21 women, 9 men, $M_{\mathrm{age}} = 18.76$, range: 18–20) were randomly assigned to either experimental ($n = 15$) or control ($n = 15$) conditions. Data from four participants were discarded because they either failed to repeat all of the training sentences aloud ($n = 3$) or had prior knowledge of a language (Persian) whose syntax matches the structures used in the semiartificial language ($n = 1$), leaving 13 participants in each group. Experimental and control groups did not differ significantly across age, sex, handedness, or number of languages (all ps > .05). All participants reported having normal or corrected-to-normal vision.

Stimuli

This section describes the three sets of materials developed and used for the elicitation of the behavioral evidence. There were two sets of exposure phase stimuli: an experimental set and a trained control set. The training instructions and training stimuli can be found in Appendix S1 of the online Supporting Information. Finally, there was a set of test-phase stimuli that all participants read during the GJT. The testing instructions and testing stimuli can be found in Appendix S2 of the online Supporting Information.

Experimental Stimuli. The experimental group was exposed to a semiartificial language consisting of English words and syntactic structures[3] based on Persian. Three syntactic structures were used to generate 96 sentences (32 sentences per structure). To create the stimuli, simple transitive English sentences were rearranged according to the three syntactic structures while still obeying within-phrase structure rules of English,[4] as in structures A, B, and C in Table 1.

Stimuli were balanced as carefully as possible within the confines of natural language. There were 5 repeating TEMPORAL PHRASEs, 20 repeating SUBJECT proper nouns (all names), nonrepeating PREPOSITIONAL PHRASEs, nonrepeating OBJECT nouns, and 48 repeated VERB phrases. Sentences were presented in random order in the training phase.

The experimental sentences contained syntactic phrase sequences that were probabilistically constrained. The transitional probabilities between syntactic phrases were either .67 or .33, depending on the transition (see Table 2). Thus, each syntactic phrase had a distributionally preferred, but not mandatory, successor in each sentence (e.g., TEMPORAL PHRASE could be followed by SUBJECT or VERB PHRASE, but was more likely to be followed by the former). Transitional probabilities were averaged over each experimental structure to calculate a mean transitional probability for each sentence type as a whole. These were as follows: Structure A = .58, Structure B = .42, Structure C = .50.[5]

Table 1 Structures used in the exposure phase of the experimental group and sentences exemplifying each

Structure	Syntactic Category Sequence	Sample Sentence
A (TSPOV)	TEMPORAL PHRASE – SUBJECT – PREPOSITIONAL PHRASE – OBJECT – VERB PHRASE	Yesterday Charlie at the supermarket milk bought.
B (TSOPV)	TEMPORAL PHRASE – SUBJECT – OBJECT – PREPOSITIONAL PHRASE –VERB PHRASE	Yesterday Charlie milk at the supermarket bought.
C (TVSPO)	TEMPORAL PHRASE – VERB PHRASE – SUBJECT – PREPOSITIONAL PHRASE – OBJECT	Yesterday bought Charlie at the supermarket milk.
D (TSPVO)	TEMPORAL PHRASE – SUBJECT – PREPOSITIONAL PHRASE –VERB PHRASE – OBJECT	Not long ago Vickie in the fridge kept a pear.
E (TSVPO)	TEMPORAL PHRASE – SUBJECT – VERB PHRASE – PREPOSITIONAL PHRASE – OBJECT	Not long ago Vickie kept in the fridge a pear.
F (TVSOP)	TEMPORAL PHRASE – VERB PHRASE – SUBJECT – OBJECT – PREPOSITIONAL PHRASE	Not long ago kept Vickie a pear in the fridge.

Note. Structures A, B, and C were grammatical items, while D, E, and F were ungrammatical items. The use of the same lexical items for each is only for illustration. Please see the Supporting Information online for all training stimuli.

Table 2 Transitional probabilities between syntactic categories in the experimental exposure stimuli

Transition	TP
TEMPORAL PHRASE – SUBJECT	0.67
TEMPORAL PHRASE – VERB PHRASE	0.33
SUBJECT – PREPOSITIONAL PHRASE	0.67
SUBJECT – OBJECT	0.33
PREPOSITIONAL PHRASE – OBJECT	0.67
PREPOSITIONAL PHRASE – VERB PHRASE	0.33
OBJECT – VERB PHRASE	0.33
OBJECT – PREPOSITIONAL PHRASE	0.33
VERB PHRASE – SUBJECT	0.33

Trained Control Stimuli. Part of the novelty of the present study is the fact that the control group also participated in a training condition (for the importance of using trained control groups, see Perruchet & Reber, 2003). The control group was trained and tested on the same stimulus sentences as far as the lexical items and the compositional semantics contained in them were concerned. However, the exposure phase sentences in the control group did not follow the three syntactic structures, but were randomized in such a way that no whole sentence structure was ever repeated (e.g., TSPOV only occurred once and temporal phrases only were sentence-initial for 1/5 of the stimuli). That is, each of the 96 sentences was presented in a different syntactic phrase order with transitional probabilities between syntactic phrases matched for all sequences. This manipulation meant that the transitional probability between any two syntactic phrases over the course of the training phase was approximately 0.25. Thus, there were no probabilistic cues to word order in the control stimuli. This provided a learning baseline that ideally isolated unforeseen task effects or mere exposure effects from the exposure phase, thus allowing them to be partialed out.

Test Materials. The 36 novel test-phase sentences consisted of the three target grammatical structures (A, B, and C from Table 1) and three ungrammatical structures (D, E, and F, from Table 1). Mostly new lexical items were used in the test phase, although some lexical items were retained from the exposure phase for readability purposes, for example, determiners, prepositions. All test sentences were constructed with two temporal phrases, six subject proper nouns (names), six prepositional phrases, six object nouns, and six verbs. Thus, the same core set of lexical items was rotated around the test stimuli. This was done to limit overacceptance or rejection of any one structure at test due to its lexical content, because each structure was drawn from the same set of lexical items. As with the exposure phase, test items were presented in random order across participants.

Procedure

Each group first participated in an exposure phase (which differed only in terms of their stimuli) and completed identical test phases.

Experimental Exposure Phase. The exposure phase was set up as a plausibility judgment task within a noncumulative self-paced reading design. Implausible sentences were defined for participants as those unlikely to happen in the real world (e.g., *Yesterday John at the store milk bought* is plausible, but

Yesterday John at the store milk sang is not). Sentences in the exposure phase were divided into four blocks of 24 sentences each. Of those 24 sentences, each structure occurred eight times, half in semantically plausible sentences and half in semantically implausible sentences. The plausibility of each sentence was determined by its final word to ensure that participants would read the entire sentence. The order of all the sentences and the four training blocks was randomized across participants.

Trained Control Exposure Phase. As with the experimental group, the control group was exposed to sentences in four randomized blocks. Likewise, half of each block consisted of semantically plausible sentences, while the other half was implausible. As stated previously, the crucial difference was that no sentence structure ever repeated in the control phase. Thus, the control group simply received four randomized blocks of randomly ordered, nonrepeating sentence structures.

General Procedure. Participants were tested individually in a quiet laboratory. They were told that they were participating in a study about meaning comprehension under time pressure (see the Supporting Information online for the full instructions). Participants were instructed to read through sentence fragments one at a time by pressing the space bar to reveal each new fragment, repeat the sentence aloud (to ensure they were paying attention), and then indicate whether the sentence depicted a scenario that was likely in the real world (i.e., make a plausibility judgment).

The experiment was administered on a PC with a 15.6” screen using SuperLab Pro 4.5. All text was black on a white background with size 18 Tahoma font. Sentences were presented in a noncumulative moving-window self-paced reading design, which required participants to press the space bar to present each fragment of a sentence. The boundaries between fragments corresponded exactly with the syntactic phrase boundaries that had been probabilistically manipulated (i.e., temporal phrase, subject, object, prepositional phrase, and verb). In other words, participants saw the constituents of a single complete syntactic phrase each time they pressed the space bar. This manipulation was intended to prevent learners from having to perform extra computations to segment the sentence.

In each exposure-phase trial, participants saw a fixation cross and then pressed the space bar to begin reading sentence fragments. They continued pressing the space bar to see each fragment until they reached the end of the sentence. They were then prompted to repeat the sentence verbatim. After doing so, they pressed the space bar again and were prompted for a plausibility judgment. After giving a plausibility judgment, participants saw another fixation cross and the next trial would begin. The average time taken to complete the exposure phase was 20 minutes.

After the exposure phase, participants were then told that the word order in the previous sentences was not random but instead had contained systematic patterns. They were instructed to read 36 new sentences, all of which would be plausible. They were told that half the new sentences would follow the same word order patterns as the previous sentences and that these should be called "grammatical." They were told that the other half of the new sentences would not conform to the same word order and should be rejected as "ungrammatical." To decrease the likelihood that they would classify sentences on the basis of their meanings, participants were reminded that all test sentences were plausible and the focus now was on word order. On average, the test phase took 5 minutes to complete.

Results for the Behavioral Experiment

Overall classification accuracy and endorsement rates on the grammatical and ungrammatical items on the GJT were taken as measures of learning for this study. Alpha levels were set to 0.05.

Overall Performance on Grammatical and Ungrammatical Items

The analysis of the grammaticality judgments (Figure 2) showed that the experimental group classified 61.89% ($SD = 16.51$) of the test items correctly and the control group 59.93% ($SD = 12.51$). The difference between the two groups was not significant, $t(24) = .35, p = .72$, indicating no overall evidence of learning. However, overall accuracy may mask differences in performance on individual structures. To further investigate this, accuracy on individual sentence structures was analyzed.

Performance on Individual Structures

In order to establish whether the experimental and control groups performed differently across the individual syntactic structures in the test phase, a 2 × 6 mixed analysis of variance (ANOVA) was conducted on participants' accuracy on the GJT with Group (2 levels: Experimental, Control) as the between-subjects factor and Structure (6 levels: A, B, C, D, E, F) as the within-subjects factor (see Figure 3). The ANOVA (with Greenhouse-Geisser correction) revealed a significant Group * Structure interaction, $F(3.06, 79.75) = 2.71$, $p = .05$, $\eta_p^2 = .17$; no effect of Structure, $F(3.06, 79.75) = 2.05$, $p = .11$;

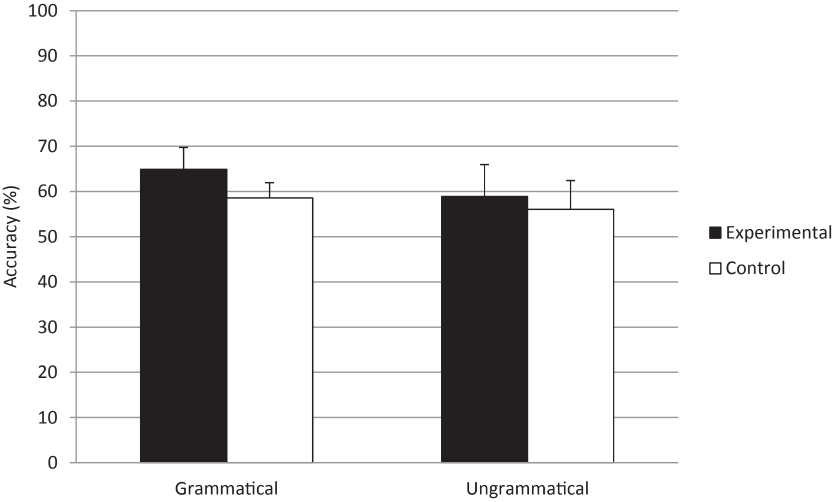


Figure 2 Mean accuracy on grammatical and ungrammatical items on the grammaticality judgment task in the Experimental and Control groups. Error bars represent standard error.

and no effect of Group, $F(1, 24) = .11$, $p = .74$. To further investigate the locus of the interaction effect, post hoc comparisons between groups were conducted. The post hocs revealed that Experimental participants significantly outperformed Controls on structure A, $t(24) = 2.44$, $p = .02$, $d = .92$, and B, $t(24) = 2.18$, $p = .03$, $d = .30$, but underperformed Controls on structure C and that difference approached significance, $t(24) = 2.03$, $p = .054$, $d = .80$. There were no between-group differences on structures D, E, or F (all ps > .34). The Experimental group outperformed the Control group on structures A and B, but underperformed Controls on structure C.

To further investigate the pattern of performance across the different syntactic structures in the GJT, Bonferroni-adjusted post hoc pairwise comparisons were conducted. The performance of the Experimental group across the different structures is reported in Table 3. Experimental participants differed in their overall endorsement of structure pairs A(TSPOV) and C (TVSPO), $p = .04$, A and E (TSVPO), $p = .01$, and A and F (TVSOP), $p = .03$. No other pairwise comparisons were significant in either group.

Overall, the results suggest some evidence of a small amount of learning in the Experimental group; however, learning was only found in structures A and B. Therefore, there is some weak evidence for statistical learning of L2 syntax

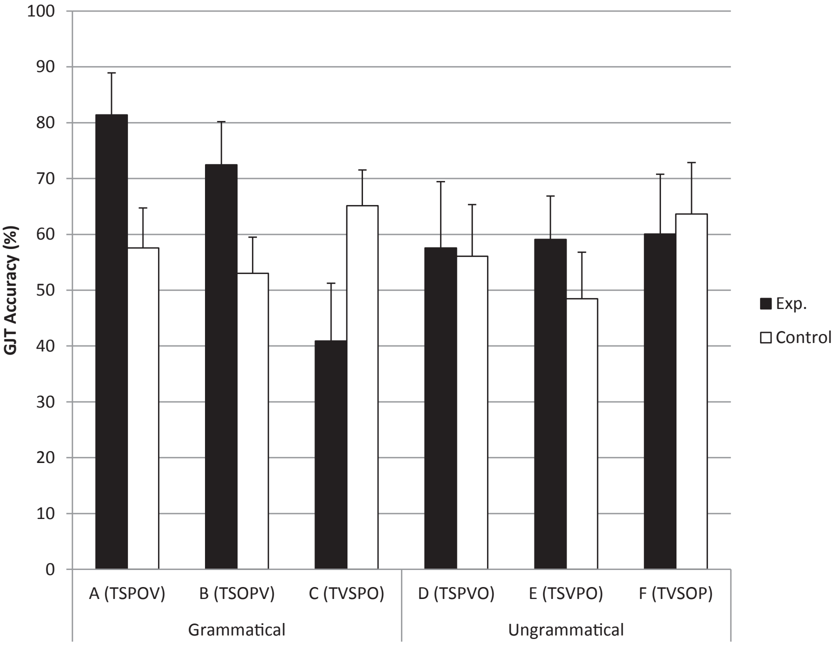


Figure 3 Mean accuracy across the difference syntactic structures on the grammaticality judgment task for the Experimental and Control groups. Error bars represent standard error.

Table 3 Stimuli coding scheme for the SRN and Parser

Structure	Coding
A	
Syntactic structure	TEMP PHRASE – SUBJ – PREP PHRASE – OBJ – VERB PHRASE
Coding	* T S P O V #
B	
Syntactic structure	TEMP PHRASE – SUBJ – OBJ – PREP PHRASE –VERB PHRASE
Coding	* T S O P V #
C	
Syntactic structure	TEMP PHRASE – VERB PHRASE – SUBJ – PREP PHRASE – OBJ
Coding	* T V S P O #

under incidental learning conditions. However, the extent to which this learning was the result of learning mechanisms that performed statistical computations must be addressed before any conclusions can be made with regard to the actual mechanisms of learning found in this experiment.

Stage 2: Comparison With Evidence From Computational Simulations

The aim of the second stage of the study, which involved computational simulations, was to investigate whether the pattern of performance for the Experimental group across the structures in the GJT is better replicated by a SRN or PARSER. If the adults in the behavioral experiment were computing predictive statistics over the syntactic categories in the semiartificial language then it would be plausible to expect the SRN to be able to replicate human performance across the different structures. However, if participants were engaging in chunk formation processes by iteratively clustering syntactic categories into larger syntactic category chunks (e.g., SPO, TSP, OV), then it would be plausible to expect PARSER to be better at capturing the human pattern of performance across the different test structures.

Method

Stimuli

Both the SRN and PARSER were trained on the same sentence structure templates as the human participants in the Experimental group. In other words, the SRN and PARSER were trained on grammatical syntactic phrase sequences. Both models were exposed to each syntactic structure 32 times in a random order, just like the Experimental group. The input coding for each sentence is shown in Table 3. In the coding scheme, letters represent syntactic phrase categories. An asterisk [*] represents the beginning of a sentence and a pound sign [#] represents the end of a sentence.[6] The lexical items of sentences were not input to either model. Instead, the coding system assumed that syntactic phrases (or their functional equivalents) are processing primitives. This was done for two reasons. First, it provided a measure of methodological comparability with previous computational studies on L2 syntax (e.g., Williams, 2010, Experiment 2). Second, as has been pointed out elsewhere (e.g., Chang et al., 2006; Williams & Kuribara, 2008), it is fairly safe to assume that adult L2 learners have recourse to some knowledge of abstract linguistic categories and apply this knowledge to their L2s.

Parameters

Following Boucher and Dienes (2003), both the SRN and PARSER were trained across a variety of parameters. This increased the likelihood that either model's fit to the human data would be due to intrinsic properties of the model rather than to idiosyncratic, specific parameters that just happened to match human behavior.

SRN. Simulated participants (matched for the number of human participants, $N = 13$) were exposed to the 96 training sentences presented in random order (i.e., 32 exposures to each syntactic structure). To test the SRN after learning, the activation of the target output node (the correct next syntactic phrase in a sequence) was recorded as a proportion of the activation of all output nodes. This is known as the Luce ratio, and it serves to measure the SRN's accuracy at predicting the next item in a sequence (for an explanation of the Luce ratio scoring procedure, see Williams & Kuribara, 2008). These values were then averaged over each test sentence to provide an index of learning in the network.

At the beginning of each simulation, the SRN contained seven localist input and output nodes (i.e., one node per syntactic phrase category plus the beginning and ending nodes) and randomly selected connection weights between the nodes. Each simulated SRN participant consisted of a single set of parameters randomly chosen from a range of values, as shown in Table 4. To ensure that a given SRN participant's performance was not due to the initial state of the network (i.e., the initial connection weights), the performance of each simulated SRN participant was calculated as the average of five individual simulations ("runs") using that SRN participant's chosen parameters. For each run, the SRN participant was reset to an initial state with connection weights selected randomly between $[-.5\ .5]$. After five complete runs, the results for that set of parameters (i.e., that SRN participant) were averaged into a score for that SRN participant. In other words, performance of each simulated participant was actually the average performance of a single set of SRN parameters (number of hidden and context units) over five independent simulations, each with random initial connection weights. This process was completed for each of the 13 simulated participants.

PARSER. As with the SRN, the simulated PARSER participants ($N = 13$) were each given a random set of parameters within the ranges listed in Table 4. Also as with the SRN, the score for each simulated PARSER participant was the average of that participant over five runs. In other words, each PARSER participant was simulated five times with the same parameters but different initial settings. To measure learning, the weights of whole syntactic structures (e.g., the unit *TSPOV# in PARSER's memory) were taken as scores. However, PARSER did not always learn an entire syntactic structure. When PARSER did not learn whole syntactic structures, chunks that did exist in PARSER's memory that had appeared in the grammatical syntactic structures had their weights summed and then averaged across runs to form a single participant's

Table 4 Parameter ranges for the SRN and PARSER

	SRN			PARSER			
	N of Hidden and Context Units	Learning Rate	Momentum	Decay	Interference	*N* of Perceptual Primitives	Initial Weight
Upper bound	20	0.4	0.0	0.10	0.020	5	1.00
Lower bound	7	0.2	0.1	0.02	0.005	1	0.75

Note. For the SRN, only variable parameters are listed, otherwise all parameters were set to default values specified in Ruh and Westermann (2009). For PARSER, only variable parameters are listed; otherwise all parameters were set to default values specified in Perruchet and Vinter (1998).

Table 5 Mean Luce ratios (SRN) and chunk weights (PARSER)

		Grammatical			Ungrammatical			Overall	
		A	B	C	D	E	F	GR	UNGR
SRN									
	M	0.528	0.502	0.512	0.421	0.404	0.381	0.514	0.402
	SD	0.124	0.103	0.115	0.068	0.054	0.065	0.114	0.061
	SE	0.034	0.028	0.032	0.018	0.014	0.018	0.032	0.017
PARSER									
	M	26.23	23.58	17.18	14.71	16.27	7.96	22.33	12.98
	SD	12.42	10.98	7.93	11.84	11.25	7.14	10.01	9.82
	SE	3.44	3.04	2.20	3.28	3.12	1.98	2.77	2.72

average performance.The coding scheme is illustrated in Appendix S3 of the online Supporting Information. This scoring procedure simply reflects the fact that participants need not have learned entire syntactic structures to endorse sentences, but could have performed exclusively on the basis of fragmentary chunk knowledge.

Results for the Computational Simulations

The results of the simulations are reported in Table 5. Simulation data from the SRN and PARSER were analyzed in order to determine the extent to which either model was able to learn the three target syntactic structures.

For the SRN, a repeated-measures ANOVA (with Greenhouse-Geisser correction) on Luce ratio scores with Structure (A, B, C, D, E, F) as within-subjects factor revealed a significant effect of Structure, $F(1.24, 14.89) = 47.70$, $p < .001$, $\eta_p{}^2 = .79$. Moreover, the SRN showed significantly higher Luce values for grammatical items than ungrammatical items, $t(12) = 7.43, p < .001$, $d = 1.22$. The results indicate that the SRN learned to discriminate grammatical from ungrammatical structures.

In PARSER, 9 out of 13 simulated participants were actually able to learn A, B, and C as whole templates. The other 4 were only partially able to do so, and required the chunk weight scoring procedure outlined above. Chunk weights were mandatorily used to compute scores for ungrammatical structures in PARSER, because PARSER could not have an ungrammatical whole structure in its memory. A repeated-measures ANOVA (with Greenhouse-Geisser correction) on chunk weights in PARSER with Structure (A, B, C, D, E, F) as within-subjects factor revealed a main effect of Structure,

Table 6 Statistical significance of Bonferroni-adjusted pairwise comparisons for human participants, the SRN, and PARSER

Structure	Structure	HUMAN Significance	SRN Significance	PARSER Significance
A	B	1.000	0.032	1.000†
	C	0.044	0.002†	0.002†
	D	0.298	0.000	0.000
	E	0.011	0.001†	0.000†
	F	0.034	0.000†	0.000†
B	C	0.159	0.643†	0.102†
	D	1.000	0.000	0.000
	E	0.272	0.001	0.002
	F	0.082	0.000	0.000
C	D	1.000	0.000	1.000†
	E	1.000	0.001	1.000†
	F	1.000	0.000	0.000
D	E	1.000	0.062	0.618†
	F	1.000	0.001	0.067†
E	F	1.000	0.411†	0.005

Note. The icon † indicates a point where the significance of human pairwise comparisons matches that of the computational model.

$F(2.43, 29.23) = 40.42$, $p < .001$, $\eta_p{}^2 = .77$. Moreover, chunk weight values were significantly larger for grammatical items than ungrammatical items, $t(12) = 25.54, p < .001, d = 0.94$. Taken together, these results demonstrate that both the SRN and PARSER were actually better at discriminating grammatical and ungrammatical items than human adults.

To investigate whether either model could better simulate the human pattern of performance in endorsement across the different items in the GJT, further analyses were conducted on the significant effect of Structure in both models. Bonferroni-adjusted pairwise comparisons were conducted on performance differences between grammatical and ungrammatical structures in the SRN and PARSER (Table 6). Recall that in the Expeirmental group, Bonferroni-adjusted pairwise comparisons showed significant differences in performance on structure pairs A-C, A-E, and A-F, while the rest of the comparisons were nonsignificant. Pairwise comparisons on the simulated performance of the SRN and PARSER showed that both models were able to simulate the significant differences in performance for structure pairs A-C, A-E, and A-F. However, PARSER was able to replicate more of the human performance patterns across the different syntactic structures than the SRN, accounting for 9/15 of the pairwise comparisons in the Experimental group, while the SRN only accounted for 5/15 of the comparisons. However, Fisher's exact test[7] shows this difference to be nonsignificant, $p = .27$; therefore, it should be treated with caution.

General Discussion and Conclusion

This study aimed to provide insights into the mechanisms of statistical learning in the context of adult learning of L2 syntax. In particular, it compared the ability of two computational models, a SRN and PARSER, to simulate adult learning of syntax in a semiartificial language paradigm under incidental conditions. The first research goal was to assess whether there would be any evidence of incidental statistical learning based on adult performance on a GJT. There was evidence of a modest learning effect in the Experimental group's performance on the GJT, but only on structures A (TSPOV) and B (TSOPV). However, contrary to expectations, the Experimental group underperformed the Control group on structure C, and the two groups were not significantly different in accuracy on ungrammatical items.

The second research goal investigated whether the human data could better be replicated by a learning mechanism that computes predictive statistics (SRN) or a learning mechanism that forms increasingly complex chunks (PARSER). PARSER was better able to replicate the pattern of human performance across the different structures in the GJT. This finding is consistent with the possibility that chunk formation processes may have been involved in the learning effects in the behavioral experiment. Importantly, the finding of an advantage for chunk-based models over connectionist models is consistent with previous research in other domains, such as word segmentation (e.g., Giroux & Rey, 2009; Perruchet & Peereman, 2004; Perruchet & Vinter, 1998) and artificial grammar learning (e.g., Boucher & Dienes, 2003). However, both the SRN and PARSER were actually better able to classify test structures than human participants, and neither model was able to fully reproduce the full range of human results. This was likely due to a combination of factors, including training the models exclusively on syntactic patterns, the exclusion of semantics, and a lack of prior knowledge of L1 in the models. Therefore, the goodness of fit of PARSER was relative to the SRN. Needless to say, other computational models, or modifications to the current models, will be necessary to fully account for the human data. Future work should also consider whether the models' robust abilities to simulate statistical learning in other domains of language and their less robust performance in the current study imply that statistical learning is less of a driving force in learning L2 syntax.

The present findings appear to contrast with previous research showing that the SRN could simulate human performance in incidental learning of semiartificial syntax (Williams, 2010; Williams & Kuribara, 2008). Why did the present study not reproduce this finding? This question necessitates further investigation, as there are a number of plausible explanations. For one, it could simply be that the SRN and PARSER simulate two different learning mechanisms that are available to humans and which may be more or less likely to operate on different types of stimuli. If this were the case, we would expect the SRN and PARSER to be able to capture human data to different extents depending on which mechanism was involved. What might cause a change in mechanism? There are many possible causes, but the tasks and procedure for the Japlish studies were very similar to those used at present. Indeed, one might suspect the semiartificial language stimuli themselves to be the chief difference. Perhaps the SRN better replicated human performance in previous work because of the more complex nature of Japlish. On this view, the relative simplicity and stability of the semiartificial language stimuli in the present study may have favored chunking mechanisms like those in PARSER. However, Williams and Kuribara (2008) reported a discrepancy between the SRN and human participants on complex long-distance scrambling Japlish structures. This suggests that stimulus complexity may not be the cause of the discrepancy. In the absence of a compelling explanation of the differences in results between the present and previous studies, it is important to keep in mind that Williams (2010) found that the SRN was best able to fit the human data when humans were trained on meaningless syntactic category analogues (96% shared variance). When the SRN was compared with performance on the actual semiartificial language learning (as was done here), the shared variance dropped to 40% and 66%. Thus, the lower degree of fit between the SRN and human behavior found here may not be so different from the results Williams obtained.

Taken together, the results from the current study suggest that chunk formation may play a role in early L2 syntactic development under incidental learning conditions. These results are consistent with previous research showing superior performance for PARSER (e.g., Giroux & Rey, 2009; Perruchet & Peereman, 2004) and other chunk-based models (e.g., Boucher & Dienes, 2003) over the SRN. Inasmuch as PARSER simulated human performance due to genuine similarities in their learning mechanisms, the present results also suggest the importance of attentional and associative learning mechanisms in the early phases of language acquisition. On this view, the attentional and associative mechanisms argued to play important roles in word segmentation (e.g., Perruchet & Vinter, 1998, 2002; Perruchet & Tillman, 2010) would also be contributing to the formation of chunks of syntactic knowledge. As such, the present results implicate general principles of attention-based associative learning and memory in early L2 syntactic development. This finding is consistent with usage-based, emergentist approaches to SLA (e.g., Ellis, 1996, 2006, 2008) and with Robinson's (1996, 1997) Fundamental Similarity Hypothesis, both of which posit the importance of attention and chunk formation in L2 development. Moreover, because PARSER simulates attentional and associative mechanisms that are implicated in various aspects of declarative memory, it is plausible to consider that the relationship between PARSER and human performance may indicate that human participants were utilizing comparable declarative memory mechanisms. If this is the case, then the present results are consistent with approaches to L2 syntax that posit a role for declarative memory in early syntactic development (e.g., the Declarative/Procedural Model, Ullman, 2004). Indeed, in a replication that extended this experiment, Hamrick (in press) found that declarative memory for syntactic information likely played a significant role in the learning process.

However, it is important to keep in mind a number of limitations that prevent strong generalizations from the present study. First, the use of the semiartificial language paradigm, while methodologically convenient, brings several limitations. Potentially the most damaging problem is that the use of English words and phrase structure recruits syntactic information that is English specific and which may cause unnatural processing or alignment problems when placed into non-English syntactic structures that would not occur in natural languages. A more concerning possibility is that the word order patterns did not match the head directionality preferences associated with English phrase structure, which was preserved in the stimuli (e.g., Greenberg, 1963, but see Dunn, Greenhill, Levinson, & Gray, 2011, for compelling counterevidence). Future ab initio learning research using natural languages is needed to assess the robustness of conclusions from the semiartificial language paradigm.

Second, the study is limited by only demonstrating a learning effect for two of the three structures in the exposure phase. Experimental participants learned structures A (TSPOV) and B (TSOPV), but performed below Controls on structure C (TVSPO). What caused this pattern of performance? Considering the aforementioned word order universals, it is possible that VSO word order created an incongruity with English phrase structure leading to poor performance on structure C. Moreover, this VSO word order may also have prompted

Controls to endorse it for being "unusual," like the randomized sentence structures they were exposed to during training. Alternatively, participants may have simply been using metalinguistic strategies that disfavored structure C. For example, in an experiment using the same semiartificial language, Hamrick (2013) demonstrated that participants were biased toward using a verb-final strategy to classify test stimuli (e.g., they recalled that verbs came at the end and then applied that simple rule in the test phase). However, when the GJT was replaced with a recognition memory task, it was found that participants performed equally well on structures A, B, and C, because the design of the recognition memory test minimized the need for metalinguistic strategies. Further research is needed to assess these and other possible explanations.

A third limitation exists in PARSER's attentional window mechanism, whose size changes randomly within the parameter range. In all likelihood this is not how human attention works, especially in adult L2 learners. L2 learners have vast prior knowledge that would, presumably, shape the size and scope of attention to the linguistic input. Moreover, L2 learners with different native languages probably have different language-specific attentional biases that they bring to the learning task (e.g., Ellis & Sagarra, 2010). This problem with PARSER speaks to a larger problem with both the computational models: namely, that neither model is generally trained on a first language before being given a second language as input. Consequently, future research is needed comparing the ability of these—and other—computational models to learn over a large L1 corpus before proceeding to L2 input.

This study is also limited by the lack of robust differences between the two computational models. The differences between the SRN and PARSER in accounting for the human data were a matter of degree, not of qualitative difference. PARSER was able to account for more of the pairwise comparisons in the human data (9/15) than the SRN (5/15), but this difference in amount of comparisons simulated was not significant. Therefore, the meaningfulness of the current findings will only be clear in the context of more research on other L2 learning data, and ongoing research with the SRN and PARSER (and other computational models) is being conducted in order to address these questions across a variety of artificial language learning paradigms. Other computational modeling using nonstatistical learning models (e.g., ACT-R; Anderson et al., 2004) is also advisable, because the present results did not show robust evidence of statistical learning.

Finally, the present study is limited in the way it coded input for both the SRN and PARSER. As has been pointed out elsewhere (e.g., Chang et al., 2006; Williams & Kuribara, 2008) it is fairly safe to assume that adult L2 learners have recourse to some knowledge of abstract syntactic categories and apply this knowledge to their L2s. Indeed, the fact that participants in the present experiment were able to generalize to new sentences implies the presence of some kind of abstract knowledge. When combined with the fact that the study was not about syntactic category induction from exemplars, it made the most sense to code the input to the SRN and PARSER as abstract syntactic categories. Moreover, because connectionist models have been shown to form their own linguistic categories (e.g., Elman, 1990), the operation of a SRN on abstract categories poses no theoretical problem. Likewise, it is not theoretically problematic for PARSER to operate on abstract categories (Perruchet, 2005; Perruchet & Gallego, 1997; Perruchet & Vinter, 2002). Indeed, the formation of increasingly complex abstract syntactic chunks is a hallmark of several theories of syntax, including construction grammar (e.g., Ellis, 2008; Goldberg, 2006; Tomasello, 2003) and simpler syntax (Culicover & Jackendoff, 2006). However, PARSER cannot abstract categories from instances like a connectionist network (it was not designed to do so) and, as such, using PARSER to simulate syntactic development will only capture the learning of syntactic sequences, not the learning of syntactic categories themselves. Future work will benefit by addressing each of these issues, perhaps by training SRNs and PARSER on surface content instead of abstract syntactic categories, or by using models that are better able to form categories from instances.

Despite its limitations, the present study offers novel advances on previous research in statistical learning and SLA. It extends previous work comparing computational models of statistical learning (e.g., Boucher & Dienes, 2003; Giroux & Rey, 2009; Perruchet & Peereman, 2004) to a novel domain: adult learning of L2 syntax. In doing so, it also extends previous research using connectionist networks, especially SRNs, to investigate L2 syntactic development (e.g., Williams, 2010; Williams & Kuribara, 2008). In conclusion, this study brings to light new and important questions regarding the mechanisms of statistical learning in L2 development. The present evidence suggests chunk formation may play a role in L2 syntactic development, but it also underscores the need for more work using the competing predictions of different computational models in order to elucidate L2 learning mechanisms.

Final revised version accepted 16 September 2013

Notes

[1] Perruchet (2005; Perruchet & Vinter, 2002) have noted that the attended information in chunks may be considered to be isomorphic with the contents of subjective

phenomenal awareness. I remain agnostic on this position at present. However, it is an important theoretical design feature that sets PARSER apart from many other computational models, which do not provide any role for awareness in learning.

[2] The semiartificial language paradigm was used for two reasons. The first was comparability with previous studies looking at the mechanisms of L2 syntactic development (e.g., Williams, 2010). The second reason was a practical choice: Using semiartificial languages reduces the duration of the experiment, since the words are usually in participants' native language, they need not be pre-trained on vocabulary.

[3] Syntactic phrase/category is used here only to denote a lexical constituent or constituents that constitute a single syntactic phrase in English. It may be that participants process argument roles or some more basic category structures rather than syntactic categories, per se.

[4] This may lead to alignment differences for combining the syntactic information in the lexicon of one language with the syntactic information from the syntax of the other. Whether or not this manifests as a problem is an empirical question that needs to be investigated.

[5] Incremental transitional probabilities were also tracked (e.g., the probability of T → S was .67, the probability of TS→P was .5, the probability of TSP→O was 1.0, and so on), but statistical analyses revealed no significant effect of incremental transitional probability on reading times (all ps $> .10$).

[6] This coding scheme was adopted for comparability with Williams's (2010; Williams & Kuribara, 2008) coding of the SRN. The same simulations were done *without* explicitly coding the beginnings and ends of the sequences this way and the results were essentially the same as those reported here.

[7] Thanks to one of the anonymous reviewers for suggesting this analysis.

References

Altmann, G. T. M. (2010). Why emergentist accounts of cognition are more theoretically constraining than structured probability accounts: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences*, *14*, 340.

Altmann, G. T. M., & Mirkovič, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 583–609.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.

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.

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science*, *27*, 807–842.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

Christiansen, M. H., & Chater, N. (2001). Finite models of infinite language: A connectionist approach to recursion. In M. H. Christiansen & N. Chater (Eds.), *Connectionist psycholinguistics* (pp. 138–176). London: Ablex.

Culicover, P., & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, *10*, 413–418.

Dienes, Z., & Altmann, G. (2003). Measuring learning using an untrained control group: Comment on R. Reber and Perruchet. *Quarterly Journal of Experimental Psychology*, *56A*, 117–123.

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*, 79–82.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*, 91–126.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 33–68). Oxford, UK: Blackwell.

Ellis, N. C. (2006). Cognitive perspectives on SLA: The associative-cognitive CREED. *AILA Review*, *19*, 100–121.

Ellis, N. C. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 372–405). London: Routledge.

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition*, *32*, 553–580.

Ellis, N.C., & Schmidt, R. (1997). Morphology and longer distance dependencies: Laboratory research illuminating the A in SLA. *Studies in Second Language Acquisition*, *19*, 145–171.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Simulating parallel activation in spoken word recognition. In M. H. Christiansen & N. Chater (Eds.), *Connectionist psycholinguistics* (pp. 76–105). London: Ablex.

Giroux, I., & Rey, A. (2009). Lexical and sublexical units in speech perception. *Cognitive Science*, *33*, 260–272.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford, UK: Oxford University Press.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

Gopnik, A., Wellman, H. M., Gelman, S. A., & Meltzoff, A. N. (2010). A computational foundation for cognitive development: Comment on Griffiths et al. and McClelland et al. *Trends in Cognitive Sciences*, *14*, 342–343.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of grammar* (pp. 73–113). Cambridge, MA: MIT Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357–364.

Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, *9*, 416–423.

Hamrick, P. (2012, October). *Associative learning supports early phases of adult L2 syntactic development: Behavioral and computational evidence*. Poster presented at the Second Language Research Forum, Pittsburgh, PA.

Hamrick, P. (2013). *Development of conscious knowledge during early incidental learning of L2 syntax*. Doctoral dissertation. Retrieved from ProQuest Dissertations and Theses database. (3558525)

Hamrick, P. (in press). Recognition memory for novel syntactic structures. *Canadian Journal of Experimental Psychology*. doi: 10.1037/cep0000002

Knowlton, B., & Squire, L. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 169–181.

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.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745–2750.

Perruchet, P. (2005). Statistical approaches to language acquisition and the self-organizing consciousness: A reversal of perspective. *Psychological Research*, *69*, 316–329.

Perruchet, P., & Gallego, J. (1997). A subjective unit formation account of implicit learning. In D. C. Berry (Ed.), *How implicit is implicit learning?* (pp. 124–161). Oxford, UK: Oxford University Press.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge. *Journal of Experimental Psychology: General*, *119*, 264–275.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, *17*, 97–119.

Perruchet, P., & Reber, R. (2003). The use of control groups in artificial grammar learning. *Quarterly Journal of Experimental Psychology*, *56A*, 97–115.

Perruchet, P., & Tillman, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science*, *34*, 255–285.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences*, *25*, 297–388.

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Reber, A. S., & Lewis, S. (1977). Toward a theory of implicit learning: The analysis of the form and structure of a body of tacit knowledge. *Cognition*, *5*, 333–361.

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures of awareness. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 249–269). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Rebuschat, P., & Williams, J. N. (2012). *Statistical learning and language acquisition*. Berlin, Germany: Mouton de Gruyter.

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*, 123–138.

Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition*, *18*, 27–67.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, *19*, 223–247.

Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*, 235–268.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*. doi:10.1002/wcs.78.

Ruh, N., & Westermann, G. (2009). OXlearn: A new MATLAB-based simulation tool for connectionist models. *Behavior Research Methods*, *41*, 1138–1143.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, *44*, 493–515.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*, 110–114.

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

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.

Servan-Schreiber, E., & Anderson, J. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592–608.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge, UK: Cambridge University Press.

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

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231–270.

Williams, J. N. (2009). Implicit learning in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 319–353). Bingley, UK: Emerald Press.

Williams, J. N. (2010). Initial incidental acquisition of word order regularities: Is it just sequence learning? *Language Learning*, *60*, 221–244.

Williams, J. N., & Kuribara, C. (2008). Comparing a nativist and emergentist approach to the initial stage of SLA: An investigation of Japanese scrambling. *Lingua*, *118*, 522–553.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Training Instructions and Stimuli
Appendix S2: Test Instructions and Stimuli
Appendix S3: Sample Scoring Procedure for PARSER