LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

Initial Incidental Acquisition of Word Order Regularities: Is It Just Sequence Learning?

John N. Williams

University of Cambridge

There is a long tradition of implicit learning research looking at learning of artificial grammars (finite-state grammars that generate meaningless letter strings). Are the associative learning processes evident in these studies at work in learning word order in natural language? To what extent do meaning and prior linguistic knowledge need to be taken into account in the natural language case? In the study reported here, incidental learning of natural language word order is compared directly to a meaningless analogue (in which the same sequential regularities underlie meaningless syllable strings). The results of both experiments are compared to connectionist (simple recurrent network) simulations. The comparisons suggest that similar associative sequence learning mechanisms underlie learning of both the natural language and its meaningless analogue (with the result that there are certain limitations to what is learned). However, to achieve this alignment, it is necessary to take into account the linguistic categories and meaning structure that the participants are likely to impose on the natural language. It is concluded that the initial incidental learning of word order can be explained in terms of associative (sequence) learning and that linguistic knowledge is engaged to the extent that it defines the categories over which statistics are computed.

Introduction

Imagine that you are in the early stages of learning a foreign language without any instruction. You have learned some vocabulary, and you are beginning to derive meaning from sentences that you hear. Let us assume that you make no conscious effort to analyze the word order in these sentences because your attention is focused on meaning. Under these circumstances, what would you learn

I am indebted to Chieko Kuribara for very helpful discussion and advice regarding this extension of our earlier work.

Correspondence concerning this article should be addressed to John N. Williams, Research Centre for English and Applied Linguistics, English Faculty Building, 9 West Road, Cambridge CB3 9DP, United Kingdom. Internet: jnw12@cam.ac.uk

about the word order regularities in the language? Would you just learn specific word sequences as multiword units (e.g., John pizza ate)? Would you learn specific structural patterns (e.g., SOV)? Or would you learn generalized grammatical rules (head-final verb position)? This is the question that the present study addresses. The answer has implications for the kind of prior grammatical knowledge that is brought to bear and for our conception of the nature of the learning process. If you only learn multiword units, then prior grammatical knowledge plays no role; if you learn specific structural patterns, then prior knowledge is at least required to identify categories of subject, object, verb, and so forth; and if you learn generalized rules, then grammatical concepts such as "head direction" may be involved, possibly suggesting a role for Universal Grammar (UG). With regard to learning mechanisms, learning multiword units might be regarded as an extension of vocabulary learning, learning structural patterns might involve domain general cognitive learning mechanisms such as chunking and sequence learning (N. C. Ellis, 1996; Lieberman, 2007), and learning generalized rules might involve UG-based learning mechanisms such as parameter setting. In general, then, the issue of the nature of what is learned can be contextualized in terms of the familiar debate between emergentist and nativist approaches to language acquisition.

There is a long tradition of research within cognitive psychology that looks at incidental learning of sequential regularities using the artificial grammar (AG) paradigm, during which participants are exposed to letter sequences generated by a finite-state grammar and afterward are asked to make grammaticality judgments. Here, too, there are debates over the nature of what is learned. Whereas some argue that participants learn the abstract structure of the grammar (Reber, 1989), others argue that they just learn chunks of letter sequence (Perruchet & Pacteau, 1990). Evidence in favor of learning abstract structure comes from observations of transfer to new letter sets (Knowlton & Squire, 1996; Matthews et al., 1989) or even modalities (Altmann, Dienes, & Goode, 1995). However, sensitivity to patterns of doubling and alternation of letters may be all that is required to support such generalization abilities (Knowlton & Squire, 1996), a sensitivity that is well attested even in infants and primates (Hauser, Weiss, & Marcus, 2002; Marcus, Vijayan, Bandi Rao, & Vishton, 1999). From this perspective, AG research provides very little evidence of learning grammars as abstract systems.

Clearly though, there are numerous problems in relating such findings directly to second language (L2) learning. Finite-state grammars do not reflect the structure of natural language and are devoid of meaning. Research of this type may therefore underestimate what is incidentally learnable when L2

learners can draw on prior linguistic knowledge and meaning. On the other hand, such experiments could be regarded as overestimating what is learnable because they utilize relatively simple systems. In a finite-state grammar there is one letter at each node. However, in a natural language, one needs to learn about the sequential structure of word classes, not actual word forms. Can the same kind of learning processes that operate at the level of letters be extrapolated to processes operating at the level of categories?

SLA researchers have explored the issue of incidental learning of word order regularities using systems based on natural languages while preserving something of the methodological rigor of the artificial grammar learning paradigm. For example, in Robinson's (2005) study of incidental learning of Samoan participants were highly accurate at accepting repetitions of trained sentences as grammatical but were poor on new grammatical and ungrammatical sentences. This suggests learning limited to actual word sequences. More encouraging evidence for learning of abstract structure comes from an earlier study (Robinson, 1996) in which, after training in a memory task, there was a degree of transfer to sentences with new lexis of a rule for forming pseudoclefts of location in English (e.g., *Where LA is is in California*). However, it is not clear whether participants learned grammatical rules as opposed to word order templates of the form *Where* N *is is* PP.

In order to test for incidental learning of generalizable rules, Williams and Kuribara (2008; henceforth W&K) examined the acquisition of Japanese scrambling. From a generative perspective, scrambling is an optional syntactic operation that moves a phrase in the direction opposite to the head direction (Saito & Fukui, 1998). So in a right-headed language like Japanese, scrambling takes place to the left. If learners have understood this principle, they will accept scrambling in contexts that they have never encountered previously or even of constituents that they have never seen scrambled in the input. In our experiment, in order to examine the early stages of learning we adopted the simple expedient of presenting sentences with Japanese syntax and case markers but English lexis—for example, *John-ga pizza-o ate* (John-NOM pizza-ACC ate, John ate the pizza). We dubbed this language Japlish. In an exposure phase, participants with no prior knowledge of Japanese were first told about the function of the case markers and then made plausibility judgments on Japlish sentences. Our question was what they would learn incidentally about the word order regularities of the language. In the exposure phase they received a majority of canonical sentences, both simple (SOV) and complex (S[SOV]V), and a minority of scrambled structures. Crucially, only scrambling of objects and adjuncts occurred in the exposure phase (e.g., OSV). In the surprise test

phase they were required to make grammaticality judgments on entirely new sentences. The results were compared with a control group that were required to perform the grammaticality judgment test without any prior exposure to Japlish (judging the likelihood that the structure would be grammatical in a language they do not know).

We found clear evidence for learning the canonical patterns received in training, even though the test items involved new lexis. Therefore, as in Robinson (1996), there was learning at the level of abstract structures (see also Cleary & Langley, 2007; Hudson Kam, 2009). However, the evidence was more mixed regarding scrambled structures, even those that had actually occurred in the input. In fact, a subset of the participants (44%) failed to reliably accept even simple scrambles that they had been trained on, showing instead a strong preference for canonical structures. We referred to these participants as "nonscramblers." Even taking just those participants who at least accepted these trained scrambles, there was no evidence of generalization to scrambling of a novel constituent (the indirect object) in simple sentences (ISOV). However, there was evidence of accepting structures in which a constituent that had been previously scrambled was scrambled in a new context (object scrambling in an embedded clause; e.g., S[OSV]V). With regard to tests of verb position, which, of course, was consistently clause-final in the input, only two out of six structures involving non-clause-final verb placement were rejected at above-chance levels (*SIVO and *IOVS). Taken together, then, these results provide evidence of learning *structural* patterns encountered in the input, although only a subset of participants learned the lower frequency ones, and no clear evidence of learning generalized notions of scrambling and verb position.

The present study is an extension of this earlier work and begins by testing whether more evidence for learning would be obtained if the amount of exposure were increased. After all, there were only 194 exposure phase sentences in the earlier experiment, and given that there were some hints of learning for some structures, it would be interesting to see if more robust effects could be obtained with increased exposure. Hence, the present Experiment 1 is essentially a repeat of the earlier experiment but with the exposure phase doubled to 388 sentences (the set of exposure sentences was simply presented twice). This manipulation also allows an alternative, cross-sectional measure of learning. In the original study we compared the experimental group with a no-exposure control group. This shows how exposure to Japlish modulates whatever initial preferences there are for structures. However, in the case in which judgments are at the same level, it is not clear whether this is because there was no effect of exposure or just that the initial preferences happen to match the preferences acquired from

exposure. Comparisons between two levels of exposure provide an arguably better test of learning because they show whether participants are on any kind of learning curve at all, regardless of the absolute level of performance.

Experiment 1

Materials

Exactly the same training and test materials were used as in W&K, in which a full rationale for their selection and construction may be found. Only a summary of the item types will be presented here.

Exposure Phase

The materials comprised sentences with English lexis but Japanese word order and case markers. The case markers were *–ga* (subject), *–o* (object), and *–ni* (indirect object). There were a total of 194 different sentences presented during the exposure phase. Of these, 138 (i.e., about 71%) were simple canonical structures of the form SV (20), SOV (36), SIOV (22) (e.g., *Pilot-ga that runway-o saw* [SOV]) and complex canonical structures of the form S[SOV]V (20) and S[SIOV]V (20) (e.g., *Barman-ga customer-ga drink-o spilled that said* (S[SOV]V)). Half of the items included at least one *wh*-word (e.g., *Student-ga dog-ni what-o offered?*). There were 20 simple items with two *wh*-words (e.g., *Bill-ga when what-o sang?*). Some of the items included an optional adjunct after S (e.g., *Horse-ga when fell? John-ga angrily Mary-ga that ring-o lost that said*.).

The remaining items involved scrambling. Most of these involved object scrambling, both in simple structures—OSV (16), OSIV (16), and complex ones—OS[SV]V (8), OS[SIV]V (8)—(e.g., *That sandwich-o John-ga ate* (OSV). *What-o Mary-ga professor-ga students-ni taught that said?* (OS[SIV]V)). Note that in the complex structures, the object is extracted from the embedded clause, a case of "long scrambling" opposed to the "short scrambling" that occurs in simple structures. There were also some simple sentences involving scrambling of an adjunct, Adj S V (8) (e.g., *When Bill-ga danced?*).

Thus, there was considerable variety in the exposure phase items in terms of the inclusion of *wh*-words and adjuncts. Only about 30% of items involved scrambling. In the majority of cases, this involved movement of the object, and in a minority of cases, this involved an adjunct. Half of the sentences of each type were semantically implausible (e.g., *Susan-ga office-in computer-o ate*.). In many items the detection of the implausibility depended on using case information (e.g., *Horse-ga Tim-ni hay-o gave.*). Hence, the plausibility judgment

task oriented the participants' attention to the meaning of the sentences and the case markers but not the order of the words as such.

Test Phase

There were 88 test structures, divided into structures that had occurred in the exposure phase, referred to as "old" structures; structures that had not occurred in the exposure phase, referred to as "new" structures; and ungrammatical structures.

The old structures comprised complex canonical structures—S[SOV]V, S[SOIV]V—short scrambling in simple sentences—OSV, OSIV—and long scrambling in complex sentences—OS[SV]V, OS[SIV]V.

With regard to the new structures, some of these involved scrambling of constituents that had been scrambled in the exposure phase, but in new contexts: scrambling of an adjunct in an SOV structure (AdjSOV) and in an OSV structure (AdjOSV), and short scrambling of the object in an embedded clause (S[OSV]V, S[OSIV]V). Other items tested scrambling of a constituent that had not been scrambled in the exposure phase: the indirect object, in simple sentences (ISOV, IOSV) and complex (IS[SOV]V) sentences. Note that the AdjOSV and IOSV items are cases of multiple scrambling. Two other sets of structures involved multiple *wh*- questions that had not occurred in the exposure set. These were *Wh*-ga *Wh*-o V (e.g., *Who-ga what-o ate*?), which follows the trained canonical SOV pattern, and *Wh-o wh-ga* V (e.g., *What-o who-ga read*?), which follows the trained OSV pattern. These will be referred to as Superiority and Superiority Reversed items, respectively. For present purposes these can be regarded as old items with novel combinations of *wh*-words.

The ungrammatical structures all violated head-final verb position. Two structures had the English head-initial position (*SVO, *SVIO), two involved right-movement of S (*IOVS, *S[OVS]V), and two involved right-movement of O (*SIVO, *S[SVO]V).

There were four sentences for each test structure, half of which contained a *wh*-word (apart from the multiple *wh*-items). They were all plausible, and repetition of lexis from the exposure phase sentences was kept to a minimum.

Procedure

The structure of exposure and test phase trials was the same as in W&K and so will only be briefly summarized here. Prior to the exposure phase, participants were informed in nontechnical terms about the function of the case markers. Japlish sentences were then presented in spoken and written form, the written form remaining on the screen until the participant made a plausibility judgment

by pressing one of two response keys. The input was staged in blocks. Block 1 contained 52 simple canonical sentences. Participants were then told that in the example sentence *Fred-ga John-ga apple-o ate that said*, Fred did the saying and John did the eating. Block 2 contained 22 complex canonical sentences. Block 3 contained a mixture of 9 complex and 18 simple canonical sentences. Participants were then told that "the words will start to come up in different orders" and in Block 4 there were 20 simple canonical, 9 complex canonical, 48 simple scrambled sentences, and 16 complex scrambled sentences. The order of trials within each block was individually randomized. Thus, the first 101 out of a total of 194 sentences had canonical word order. For the 388 exposure group, this first pass through the exposure sentences was followed by a second pass, but this time there was no division into blocks and all of the sentences were simply presented in an individually determined random order.

Prior to the test phase, participants were told that "the grammar of Japlish allows a variety of word orders. All of the sentences you have had up until now were grammatical in Japlish. We are now interested in your intuitions about which sentences you think are likely to be grammatical, and which sentences you think are not likely to be grammatical, in Japlish" (the no-exposure control group were told to judge "which word order patterns seem to you to be more likely to be grammatical in a language that you do not actually know"). Sentences were again presented auditorily and visually, but this time the visual form disappeared as soon as the auditory sentence ended. This was to encourage immediate responding. Participants made a grammaticality judgment using one of two response keys and then had the option of repeating the trial if they so wished. Each sentence was preceded by a schematic representation of its meaning (see W&K for an example) in order to facilitate meaning interpretation and help the participants focus on word order.

Participants

In W&K there were 16 participants in the no-exposure group and 25 in the 194 exposure group. In the present experiment there were an additional 25 participants in the 388 exposure group. All participants were native English speaking university undergraduates and postgraduates with no prior knowledge of Japanese.

Results

For the purposes of analysis, the test structures were grouped as follows (the multiple *wh*-items have been excluded from this analysis because they are neither exactly like old canonical sentences nor are they new scrambles):

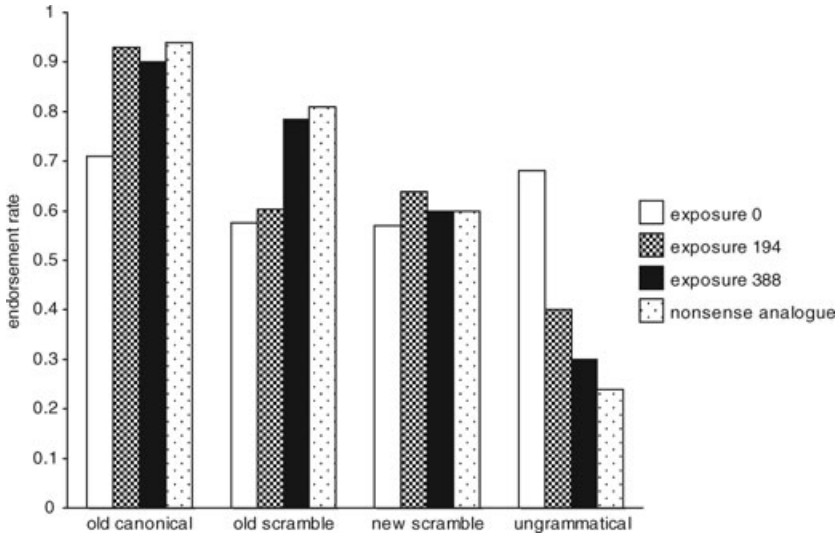

Figure 1 Endorsement rates for groups of test structures in W&K (0 and 194 exposure), Experiment 1 (388 exposure), and Experiment 2 (nonsense analogue).

Old canonical:	S[SOV]V, S[SIOV]V
Old scrambles:	OSV, OSIV, OS[SV]V, OS[SIV]V
New scrambles:	ISOV, AdjSOV, AdjOSV, S[OSV]V, S[OSIV]V, IS[SOV]V
Ungrammatical:	*SVO, *SVIO, *IOVS, *SIVO, *S[OVS]V, *S[SVO]V.

The mean endorsement rates for these groups of structures in the original W&K data and the present experiment, where there was twice the exposure, are shown in Figure 1. For each group of structures, a univariate ANOVA was performed to test for a main effect of exposure group, followed up by post hoc Sheffe tests to explore individual comparisons between participant groups.

Old Canonical Structures

There was a main effect of exposure group, $F(2, 63) = 51.1$, $p < .001$. Post hoc comparisons showed that the no-exposure group differed significantly from both the 194 and 388 exposure groups ($p < .001$ in both cases) but that the latter did not differ from each other.

Old Scrambles

There was a main effect of exposure group, $F(2, 63) = 6.32$, $p < .05$. The 388 exposure group differed from both the no-exposure group ($p < .05$) and the

194 exposure group ($p < .05$), but the latter two groups did not differ from each other. Thus, the initial exposure in the 194 group was not sufficient to alter the preexperiment level of preference for these items. However, there was a significant increase in acceptance of these items after more exposure.

New Scrambles

There was no effect of group ($F < 1.0$). Item-by-item analyses using independent sample *t*-tests showed that none of the structures in this group showed a significant difference between the 194 and 388 exposure groups, with differences (388 exposure minus 194 exposure) ranging from 0.04 (IS[SOV]V) to –0.17 (AdjOSV).

Ungrammatical Structures

There was an effect of group, $F(2, 63) = 12.75$, $p < .001$. The no-exposure group differed significantly from both the 194 exposure group ($p < .01$) and the 388 exposure group ($p < .001$). The 194 and 388 exposure groups did not differ from each other. Thus, there was a large decrease in endorsement rate for these structures after 194 exposure sentences but little further decrease with twice the exposure.

Discussion

The results show clear evidence of the impact of exposure on grammaticality judgments, indicating incidental learning of word order patterns. There was a significant increase in endorsement rates for complex canonical structures after 194 exposure sentences but no further increase after 388, possibly because of a ceiling effect. The rapid learning of these structures is not surprising given their high frequency in the input. There was also a significant increase in endorsement of old scrambles after 388 exposure sentences but not after 194. Because the test sentences had novel content, it is clear that whatever learning was occurring was not at the level of literal word sequences but rather at the level of grammatical categories. This is consistent with previous research showing that incidental learning of novel word order patterns generalizes to sentences with new lexis (Cleary & Langley, 2007; Hudson Kam, 2009; Robinson, 1996).

Despite there being good evidence for ultimate learning of structures that had occurred in the input, even at low relative frequencies, there was no evidence of generalization of scrambling to new grammatical structures, even after extensive exposure. This is particularly surprising for the S[OS(I)V]V structures because these involve a simple scrambled structure that was encountered in the exposure phase, in the form of OS(I)V, but in an embedded rather than

a main clause in the test. Thus, not only was there no evidence for learning a generalized rule of scrambling but also no evidence of generalizing clause-level word order patterns to new contexts.

Overall endorsement rates on ungrammatical structures also failed to show a significant drop with increasing exposure, pointing to a failure to learn generalized rules concerning verb position. This may seem surprising given that every clause encountered in the exposure phase ended with a verb. However, it is not the case that participants learned nothing about verb position. Endorsement rates in the 194 group were significantly lower than for the no-exposure controls. Why then was there no further significant drop with double the exposure?

The failure to find improvements in performance for the new scrambles and ungrammatical items with increasing exposure can be explained by assuming that participants do not base their judgments on rules but rather on level of similarity to structures encountered in the exposure phase. A similarity-based strategy does not lead to categorical rejection or endorsement of item types but rather to variable judgments that, averaged over tokens and participants, will reflect that item type's level of similarity to the exposure data. The important point about such a judgment strategy is that it is based on input statistics, and no matter how many times the training set is presented, these input statistics do not change. So judgments on new items (whether new scrambles or ungrammatical items) remain stable with increasing exposure. Thus, there can be a significant difference between judgments on ungrammatical items in the 194 and no-exposure groups, as only the 194 exposure group's judgments are based on input statistics. However, with increasing exposure there is relatively little change to those judgments, apart from any changes due to more accurate computation of input statistics, at least so long as a similarity-based metric is being used.

Of course the above argument applies to old structures as well, so why did increased exposure lead to performance improvements in this case? One possibility is that decisions come to be based on recognition of familiar structural patterns as these become more firmly established in memory with greater exposure. An analysis in terms of scrambler and nonscrambler participants is relevant here. Recall that in W&K, 11 out of 25 (44%) exposure group participants were classified as "nonscramblers" because they failed to accept even simple trained scrambles. In the present 388 exposure group there were still some participants who behaved in this way, although the proportion dropped to 7 out of 25 (28%). Thus, some participants persist in ignoring structures that are relatively infrequent in the input, regardless of their actual frequency. A similar tendency has previously been observed not only in child learners (Hudson Kam

& Newport, 2005) but also in adult learners for whom there are high levels of variability (Hudson Kam & Newport, 2009). Whatever the explanation of these individual differences, there is a suggestion here that recognition of trained patterns involves processes that go beyond input statistics.

The question now is, assuming that participants were not learning rules, how can we characterize the nature of the knowledge that was used to make judgments on new items? To say that performance on new items was driven by input statistics raises the question of what kind of statistic was computed and over what kinds of representation. The hypothesis that motivated Experiment 2 was that participants learned about the sequential structure of the exposure items and that training sentences were encoded in terms of sequences of grammatical categories, possibly aided by the fact that some of these categories (subject, object, and indirect object) were marked by case markers. This is to say that additional linguistic knowledge, beyond that required to recognize the relevant categories, did not contribute to the learning process. In Experiment 2 this hypothesis was tested by creating a nonsense analogue of Japlish in which lexical items were replaced with classes of nonsense words corresponding to grammatical categories. For example, the Japlish sentence *Horse-ni farmer-ga hay-o gave* was translated as the syllable sequence *to-ni so-ga pa-o ku*, where *to*, *so*, *pa*, and *ku* were members of distinct nonword classes corresponding to the grammatical categories indirect object, subject, object, and verb, respectively. Thus, sequences of grammatical categories were translated into sequences of form-level categories that could not be identified with grammatical functions or meanings. Yet the sequencing of the categories remained the same as in Experiment 1.

Research on statistical learning has revealed remarkable human and indeed primate abilities to learn about the sequential structure of meaningless syllable sequences. Adults, as well as infants have been shown to be sensitive to the probabilities of which syllables are likely to follow which and to use this information to discover where potential word boundaries might be located (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin 1996). Other studies have shown how information about the sequential probabilities of nonsense words can be used to help discover underlying phrase structure (Saffran, 2001). The purpose of Experiment 2, therefore, was essentially to see to what extent learning Japlish was driven by similar sequential learning processes.

Experiment 2

Materials

The nonsense analogue was based on the 388 exposure condition of Experiment 1.

First, sets of nonsense syllables were created for each grammatical category as follows: Subject, *si/se/sa/so*; Object, *pi/pe/pa/po*; Indirect Object, *te/ta/to/tu*; Verb, *ki/ka/ko/ku*; Adjunct, *di/da/du/de*; Adverb, *ri/ra/ro/re*. The use of a common onset for each class was intended to facilitate recognition of items as belonging to classes. Because individual *wh*-words are unique, but phonologically similar items in Japlish they were also unique, but phonologically similar, items in the analogue: What, *fu*; Who, *fe*; When, *fa*. The complementizer *that* was replaced with *me*. Next, each sentence in the exposure set was translated into a nonsense analogue version by randomly substituting lexical items for nonsense syllables of the appropriate class, retaining case markers. So, for example, the Japlish sentence *John-ga quickly beer-o drank* became *sa-ga re pe-o ku*, *John-ga Mary-ga friend-ni present-o gave that said* became *se-ga so-ga tu-ni pi-o ko me ka*, and *Tim-ga who-o hit?* became *si-ga fe-o ku?* Materials for the test phase were constructed in the same way, ensuring that no string corresponded exactly to a string in the exposure set. The items were recorded by the author, preserving the intonation patterns used in the original Japlish.

Procedure

The procedure was analogous to that used for the 388 exposure group in Experiment 1. Strings were presented auditorily and visually, with visual presentation synchronized with the audio. Instead of making plausibility judgments on exposure phase sentences, participants now performed a probe recognition task. Each string was followed by a visually presented nonsense syllable (with case marker where appropriate) and participants were required to judge whether it exactly matched the string that they had just received. Mismatches could involve a whole element that was not in the string or just part of an element (e.g., correct syllable but wrong case marker). Half of the trials contained a mismatching probe. In the test phase, participants were first informed that the strings they had just been exposed to "obeyed a set of rules that determined the possible ordering of the elements" (where an element was defined as *ra, so-ga, tu-ni*, etc.). They were then asked to judge whether the following strings were "consistent with the same set of rules that generated the strings in the first task." They were told that no string would be identical to a string received previously and that the elements would be the same as those in the exposure strings, there being no recombination of parts of elements. Test phase items were presented in the same way as in Experiment 1.

Participants

A total of 17 graduate and postgraduate students participated, none of whom knew Japanese.

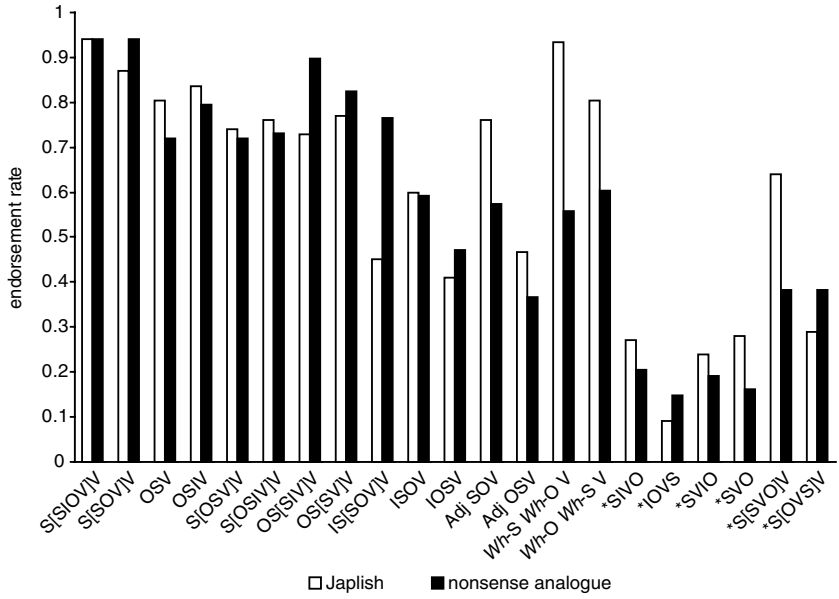

Figure 2 Endorsement rates for each item type in Experiment 1 (Japlish) and Experiment 2 (nonsense analogue).

Results and Discussion

Figure 1 shows the mean endorsement rate for the old canonical, old scramble, new scramble, and ungrammatical sets of items. There is a close correspondence between the nonsense analogue and the 388 Japlish condition. There were no significant differences between the groups on any sets of items.

For a finer grained analysis, the mean endorsement rate was calculated for each type of test item ($n = 21$). The Pearson correlation between the 388 Japlish and nonsense analogue data sets was $r = .832$, $p < .001$, meaning that 69% of the variability in the artificial analogue data was accounted for by Japlish endorsement rates. The close correspondence between the two datasets suggests that similar sequence learning processes were operating in the two experiments.

Figure 2 shows the mean endorsement rate for each type of test structure in Experiments 1 and 2. Although there is a close correspondence between absolute levels of performance for many test structures, there are also some structures that show large divergences. Between-groups differences were evaluated for each structure using the Z-ratio for the difference between independent

proportions. A nonparametric test was chosen because there were only four observations per participant per structure, and for a number of critical structures, variance was not homogeneous between groups.

Only two structures showed a significantly higher endorsement in the nonsense version than Japlish, these being OS[SIV]V ($Z = 2.64$, $p < .01$) and IS[SOV]V ($Z = 4.05$, $p < .001$). Extraction from an embedded clause produces an unusual mapping between form and meaning, and it is not unreasonable to assume that this leads to a sense of ungrammaticality. Clearly, this is not a relevant factor in the analogue. Note that the complexity engendered by long scrambling does not appear to adversely affect learning. Trained long scrambles show, if anything, a greater improvement between the 194 and 388 exposure groups than the trained short scrambles (differences of 0.22 and 0.14 respectively).

Next, consider the multiple *wh*-items *Wh*-S *Wh*-O V, and *Wh*-O *Wh*-S V. Here, the endorsement rate was higher in Japlish than the analogue ($Z = 5.91$, $p < .001$, and $Z = 2.79$, $p < .01$ respectively). Note that in Japlish the endorsement rate for *Wh*-S *Wh*-O V is as high as other canonical structures, and the rate for *Wh*-O *Wh*-S V is exactly the same as OSV. These multiple *wh*-structures are simply treated as SOV and OSV items. Clearly, participants understand the functional equivalence of *wh*-words and nouns in Japlish. However, because in the nonsense analogue these two classes of word correspond to phonologically different syllable classes, there is no reason to regard them as equivalent, hence the lower acceptance rates for novel multiple *wh*-items. This is an example of how linguistic knowledge influences learning of Japlish.

Notice also that AdjSOV is endorsed more highly in Japlish than the analogue ($Z = 2.54$, $p < .05$), although in this case not as highly as canonical structures. If participants regard an adjunct as an optional element that is freer to move than other constituents, then they would regard this structure as an acceptable variant of the canonical SOV with which they are highly familiar. Endorsement rates would be inflated relative to the analogue for which the adjunct syllable class has no special status. Again, linguistic knowledge influences Japlish.

The only ungrammatical structure to show a divergence was *S[SVO]V, for which endorsement rates were higher than in the analogue ($Z = 3.36$, $p < .001$). Indeed, they were higher than for any other grammatical structure. It is tempting to attribute this to first-language (L1) syntactic knowledge. Although endorsement rates for the simple sentences with English word orders *SVO and *SVIO were relatively low, an SVO structure that is embedded within a complex sentence may be more difficult to detect. This item could therefore

reflect L1 transfer. However, it could also be that what makes this item seem so grammatical is that it has the characteristic S[S .. (*–ga –ga*) beginning of canonical complex structures. The present data do not allow us to adjudicate between these alternatives, but see the connectionist modeling below for further evidence on this point.

Apart from these differences in the behavior of a minority of items, the overriding impression from Experiment 2 is that similar learning mechanisms are operative in Japlish and the nonsense analogue. The processes involved in deriving a meaning for Japlish only appeared to influence the long scrambles. Linguistic knowledge contributed insofar as it allowed an identification between *wh*- and non-*wh*-phrases in Japlish and possibly a more flexible approach to the positioning of adjuncts. The hypothesis suggested by these observations is that linguistic knowledge influences Japlish by determining the nature of the categories over which sequence learning processes operate. In the analogue, the categories were simply defined by phonological form, but in Japlish, the categories were of a more abstract grammatical kind, and as such, came with ancillary assumptions about their behavior. Thus, the underlying learning mechanism could be the same in the two cases, the only difference would be the nature of the representations over which the mechanism operates. The only example of a more specific L1 influence at the level of syntactic knowledge was the high endorsement of *S[SVO]V, but, as we will see, L1 transfer may not provide the correct explanation for this effect.

Connectionist Simulations

So far it has simply been assumed that a form of sequence learning is involved in both learning Japlish and the nonsense analogue. However, there are other models of learning that would not necessarily be characterized as sequential, such as chunking models that have been successfully applied to artificial grammar learning (Cleeremans & Dienes, 2008). How, then, can we be sure that sequence learning is involved?

The approach adopted here is to examine the fit between the human data and a computational model of sequence learning. Connectionist simple recurrent networks (SRNs) have been successfully applied to learning sequential information in the domains of lexical segmentation (Christiansen, Allen, & Seidenberg, 1998; Elman, 1990), artificial grammar and serial reaction time tasks (Cleeremans & McClelland, 1991; Kinder & Shanks, 2001), and natural language syntax (Chang, Dell, & Bock, 2006). Like other connectionist models, these map input representations onto output representations through a layer

of hidden units. In a sequential learning task, the network is essentially trained to use the current event represented on the input layer to predict the next event as a representation on the output layer. The network also has a "memory" for preceding events—hidden unit activations from preceding cycles of processing are copied onto "context" units. This pattern is combined with the representation of the next event and projected onto the hidden units so that the network's predictions are based not just on the event but also a memory of preceding events. This form of memory appears to degrade over time in a humanlike way (Cleeremans & McClelland, 1991).

In W&K we described an SRN model of the Japlish 194 exposure group data. Exposure phase sentences were presented to the network coded as grammatical categories. So, for example, for the exposure item *John-ga book-o read*, the input unit for S was activated and the network was told that the correct prediction was O. Connection weights were adjusted according to the degree of error on the output units, and the hidden unit activation pattern produced by S was copied to the context layer. Then the next element O was activated on the input layer and projected onto the hidden units along with the pattern from the context units (the memory of S). The network was told that the correct prediction was V, weights were adjusted according to the error on the output units, and the hidden unit activations were copied to the context units (the memory of S and O). Then V was presented, and the network was taught to predict the end of the sentence. This procedure was repeated for a total of either 50 or 250 cycles through the exposure set sentences. In the test phase, test sentences coded as grammatical categories were presented. For example, for a test item such as OSV, the unit for O was activated on the input and the strength of the correct prediction, S, was calculated on the output. The hidden unit activation was copied onto the context layer and combined with the next element, S, and the strength of the correct prediction V was calculated. Then V was presented and the strength of the correct prediction—End of sentence—was calculated. Finally, the average strength of the correct predictions over the whole sentence was calculated to give a measure of how well the network was able to predict each successive element of the sentence. The average prediction strength for each structural type in the test was then compared to the human data using a linear regression model. There was a good fit to the human data, although the results will not be discussed here because the results of a modified simulation will be presented below.

The first question to be addressed by the present simulations was whether an SRN would provide a good fit to the nonsense analogue data from Experiment 2. This is a simpler learning situation than that for Japlish, not affected by linguistic

or semantic factors, and so if an SRN cannot model these data, it is unlikely to be successful at modeling Japlish.

One input unit was assigned to each class of nonsense syllable (i.e., one unit for *si*/*se*/*sa*/*so*, another for *pi*/*pe*/*pa*/*po*, and so on), one unit for each case marker (–ga, –o, –ni), units for beginning and end markers of strings, and one unit to encode blanks between strings, resulting in a total of 16 input units. The same coding was applied over the 16 output units, and there were 9 hidden and 9 context units. The simulation was run using Tlearn (Plunkett & Elman, 1997) and averages taken over 10 simulations, each with different randomly assigned initial weights (the learning rate was 0.1 and the momentum was 0.0). For a network trained over 50 cycles the fit to the human judgment data from Experiment 2 was very good, with an r^2 value of .888, $p < .001$. A separate set of simulations with training increased to 250 cycles produced a remarkably high fit of $r^2 = .959$, $p < .001$ (see Figure 3). In other words, the network was able to account for nearly 96% of the variability between different test items in the human data.

Learning the nonsense analogue therefore appears to be accounted for remarkably well by the kind of sequential learning mechanisms instantiated by

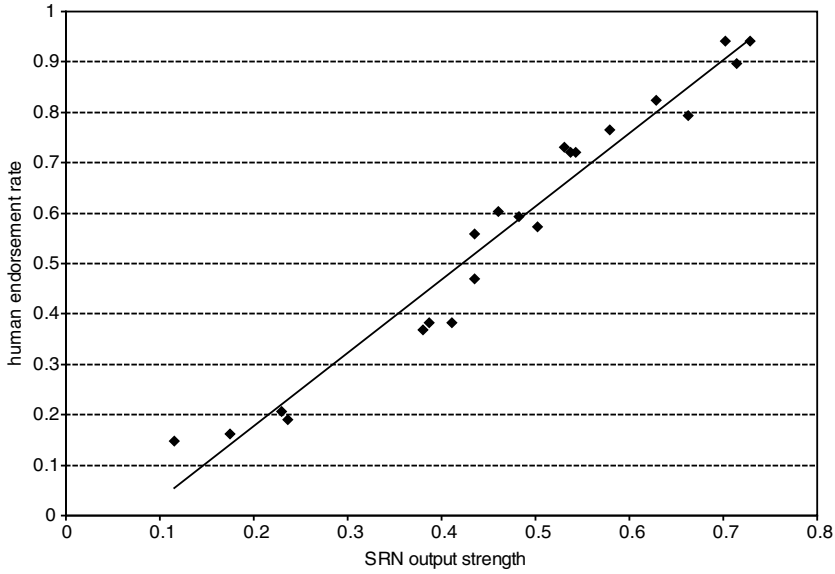

Figure 3 SRN output strength plotted against human endorsement rates in Experiment 2.

the SRN. Note that there is no claim here that the SRN actually models learning processes in the brain. The claim is merely that the SRN is computing similar information as the brain. It will be assumed that the nature of that computation can be broadly described as contingency learning (N. C. Ellis, 2006)—that is, learning the context-dependent contingencies between events in the sequences.

The next question is, How well will the above simulation account for the Japlish data? Given that the nonsense stimuli are translations of the Japlish stimuli, all that is necessary is to fit the SRN outputs from the above simulation to the Japlish data instead of the analogue data. This resulted in r^2s of only .396, $p < .01$, for the 194 exposure group and .66, $p < .001$, for the 388 exposure group. Why was there this reduction in the fit of the model?

The answer is simple. The coding used for the nonsense analogue is not suitable for Japlish because it does not take into account the potential contributions of linguistic knowledge and meaning to Japlish that were noted in Experiment 2. A number of changes were therefore made to the coding scheme. First, the coding reflected underlying grammatical categories rather than forms. For example, *John-ga* was coded simply by one subject unit, and book-o was coded by one object unit. Second, for *wh*-words, a *wh*-unit was activated at the same time as the S, O, I, or adjunct unit; for example, *what*-o was encoded as *wh* + O. This preserves the underlying similarity of *wh*- and non-*wh*-arguments and adjuncts. Finally, a cue to embedding was provided by activating an additional "matrix" unit, M, to the S and V inputs when they were part of the matrix clause of a complex sentence. For example, S[SOV]V would be encoded as S+M S O V V + M. Note that this unit was not added for simple sentences (e.g., SOV). Thus, the similarity between a simple SOV and an embedded SOV structure was preserved (this is in line with the way the way the participants were instructed). The new coding scheme resulted in 12 input units. The network had nine hidden units and was trained for either 50 or 250 cycles through the exposure set, with 10 runs per simulation.

The fit of the linear regression was better than for when the nonsense analogue coding was applied to these data. For the 50 cycle simulation, the r^2 was .563 when fitted to the 194 group data ($p < .001$) and .785, when fitted to the 388 group ($p < .001$). For the 250-cycle simulations the fit improved to $r^2 = .570$ and $r^2 = .826$, respectively. Figure 4 shows the data for the 250-cycle simulations. A striking feature of the 194 simulation is that the three points for which the network predicts a higher endorsement rate than is obtained (a, b, and c in Figure 4a) correspond to long scrambles—OS[SV]V, OS[SIV]V, and IS[SOV]V. When the participants receive more training, the trained long scrambles improve (points a and b in Figure 4b), but IS[SOV]V

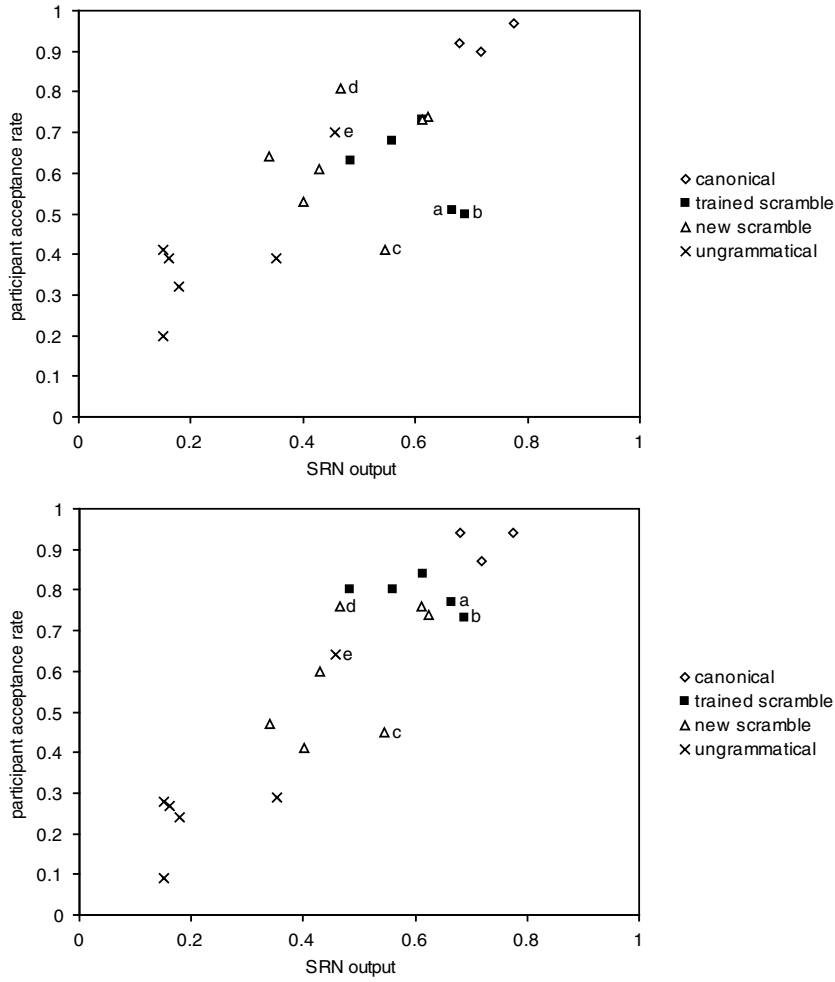

Figure 4 (a) SRN output plotted against the 194 exposure group data from Experiment 1; **(b)** SRN output plotted against the 388 exposure group data from Experiment 1.

remains problematic (point c in Figure 4b). We have already seen that these structures do not behave as predicted by the nonsense analogue, suggesting problems interpreting test sentences that involve extraction from an embedded clause. Clearly, this is beyond the scope of the network to simulate even when embedding is encoded in the input, lending further support to the idea that

it reflects the influence of processing difficulty when making grammaticality judgments. It is also interesting that AdjSOV is also a poor fit to the model (point d in Figure 4). This structure did not behave as predicted by the nonsense analogue, and because no change to the coding scheme was made in relation to adjuncts, it is not surprising that it does not fit the simulations either. With the long scrambles and AdjSOV removed from the analysis, the fit to the 250-cycle simulation increases to $r^2 = .889$ for the 194 exposure group and $r^2 = .902$ for the 388 exposure group, levels of fit that approach those obtained for the nonsense analogue simulation.

Finally, note that performance on *S[SVO]V is as predicted by the model (point e in Figure 4). This suggests that the high endorsement rate for this structure is not a reflection of L1 influence but rather simply its level of similarity to trained structures. The S[S pattern is characteristic of canonical complex sentences, and SV and OV are fragments of frequent canonical structures received in the exposure phase.

Conclusion

The question posed at the beginning of this article was what kind of knowledge of word order regularities would be incidentally acquired on initial exposure to a novel language. In Experiment 1, participants learned abstract structures underlying sentences to which they had been exposed. Transfer to test items with new lexis appears to be trivial in the case of natural language, whereas in the artificial grammar learning paradigm, transfer to new letter sets or modalities generally results in very weak effects (Altmann et al., 1995; Matthews et al., 1989). The reason is presumably that in natural language, sentences are encoded in memory in terms of grammatical categories as well as, presumably, word forms (although the latter was not tested here).

However, just because participants learned abstract structural representations does not mean that they learned rules. Experiment 1 provided no evidence of learning generalizations relating to verb position and scrambling even after doubling the exposure. Similarly, in artificial grammar learning research when truly abstract rule systems are targeted, no implicit learning is obtained (Perruchet, 1994; Shanks, Johnstone, & Staggs, 1997). This could point to limitations of the kind of associative learning processes that are operative in the two cases (Williams, 2009). Experiment 2 and the connectionist simulations did indeed suggest that similar sequence learning mechanisms were operative in Japlish and the nonsense analogue.

However, just because similar learning processes are operative in the natural language and its nonsense analogue does not mean that the learning outcomes have to be identical. This is because the representations over which the learning processes operate are different (see Patel, 2003, for a similar argument in relation to language and music). Here, differences between the two experiments and the simulations could be readily attributed to the contribution of linguistic knowledge to the way in which Japlish was encoded (i.e., in terms of linguistic categories such as subject, object, *wh*-word, adjunct). Assuming an appreciation of embedding also improved the fit of the connectionist model to Japlish. However, it was not necessary to abandon the assumption that the same underlying sequence learning processes were operative in Japlish and the analogue.

With regard to generality of the results, first it is obvious that an emphasis on learning sequences of grammatical categories assumes prior knowledge of what those categories are. Therefore, the present conclusions are only relevant to adult SLA, where it is reasonable to assume that L2 words inherit syntactic information from L1 translation equivalents and where notions of subject, object, and so forth can be transferred from the L1. Evidence for cross-language syntactic (Salamoura & Williams, 2007) and gender priming (Salamoura & Williams, 2008) certainly suggests commonality of lexical-syntactic information even in quite advanced L2 learners, and so it is reasonable to assume the same for early learning.

Second, it has to be kept in mind that both Japlish and the nonsense analogue contained case markers. It is not clear to what extent these cues helped participants track the sequence of grammatical categories, even though there were other unmarked categories as well. How well can category sequences be learned without case markers? We have also conducted experiments similar to the present Experiment 1 using German word order patterns, examining the acquisition of verb position rules (Rebuschat, 2009; Rebuschat & Williams, 2006). As here there was good evidence for incidental learning of trained patterns, as evident in performance on test sentences with new lexis, but there was no evidence of actually learning the verb position rules. However, the extent to which sequence learning processes underlie learning of this system remains unclear at present.

Third, to what extent are these results specific to early learning? In W&K we argued that in fact the limited scope of incidental learning in these experiments is indicative of the limitations of associative learning (at least of the form instantiated in SRNs) rather than limitations of exposure. This is because no matter how long an SRN is trained on Japlish, it still shows improvements

only on old structures, not new scrambles or ungrammatical structures. The implication is that if L2 learners acquire generalized rules, this must be through some other learning process. For example, it has been suggested that in L2 learners, the ability to reliably reject ungrammatical structures is associated with explicit knowledge (R. Ellis, 2005). On the other hand, if such behavior could be shown instead to be associated with implicit knowledge, then it would be necessary to appeal to other, possibly UG-guided learning mechanisms. Thus, the argument is not that generalized word order rules are ultimately unlearnable, only that they would be if learners were to persist with the simple sequence learning mechanisms evident in early learning.

References

Altmann, G. T. M., Dienes, Z., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 899–912.

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

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*, 221–268.

Cleary, A. M., & Langley, M. M. (2007). Retention of structure underlying sentences. *Language and Cognitive Processes, 22*, 614–628.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *Cambridge handbook of computational psychology* (pp. 396–421). Cambridge: Cambridge University Press.

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

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. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*, 1–24.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141–172.

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

Hauser, M. D., Weiss, D. J., & Marcus, G. F. (2002). Rule learning by cotton-top tamarins. *Cognition, 86*, B15–B22.

Hudson Kam, C. L. (2009). More than words: Adults learn probabilities over categories and relationships between them. *Language Learning and Development, 5*, 115–145.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*, 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59*, 30–66.

Kinder, A., & Shanks, D. R. (2001). Amnesia and the declarative/nondeclarative distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience, 13*, 648–669.

Knowlton, B. J., & Squire, L. R. (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.

Lieberman, P. (2007). The evolution of human speech: Its anatomical and neural bases. *Current Anthropology, 48*, 39–66.

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

Matthews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J.-R., & Druhan, B. (1989). The role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1083–1100.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience, 6*, 674–681.

Perruchet, P. (1994). Learning from complex rule-governed environments: On the proper functions of non-conscious and conscious processes. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and unconscious information processing* (pp. 811–835). Cambridge, MA: MIT Press.

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

Plunkett, K., & Elman, J. L. (1997). *Exercises in rethinking innateness: A handbook for connectionist simulations*. Cambridge, MA: MIT Press.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*, 219–235.

Rebuschat, P. (2009). *Implicit learning of natural language syntax*. Unpublished doctoral dissertation, University of Cambridge, Cambridge.

Rebuschat, P., & Williams, J. N. (2006). Dissociating implicit and explicit learning of natural language syntax. In R. Sun & N. Miyake (Eds.), *Proceedings of the twenty-eighth annual meeting of the Cognitive Science Society* (p. 2594). Mahwah, NJ: Erlbaum.

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. (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 & Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition, 27*, 235–268.

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

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

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

Saito, M., & Fukui, N. (1998). Order in phrase structure and movement. *Linguistic Inquiry, 29*, 439–474.

Salamoura, A., & Williams, J. N. (2007). Processing verb argument structure across languages: Evidence for shared representations in the bilingual lexicon. *Applied Psycholinguistics, 28*, 627–660.

Salamoura, A., & Williams, J. N. (2008). The representation of grammatical gender in the bilingual lexicon: Evidence from Greek and German. *Bilingualism: Language and Cognition, 10*, 257–275.

Shanks, D. R., Johnstone, T., & Staggs, L. (1997). Abstraction processes in artificial grammar learning. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 50*, 216–252.

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 Publishing.

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.