

Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Mikael Xie

April 12, 2022

Artificial Language Learning:
The Concurrent Acquisition of Word Order and Semantics

by

Mikael Xie

Ben Wilson
Adviser

Psychology

Ben Wilson
Adviser

Donna Maney
Committee Member

Philip Wolff
Committee Member

Sarah Fankhauser
Committee Member

2022

Artificial Language Learning:
The Concurrent Acquisition of Word Order and Semantics

By

Mikael Xie

Ben Wilson

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Psychology

2022

Abstract

Artificial Language Learning: The Concurrent Acquisition of Word Order and Semantics By Mikael Xie

Artificial grammar learning (AGL) paradigms are used to test the mechanisms of language learning among adults, children, and even infants. In AGL studies, structured sequences of stimuli are taught to participants to assess how people learn syntactical rules. In order to isolate grammar learning, the traditional AGL method is to use meaningless nonsense stimuli, such as fake words or arbitrary images; however, real-world languages do not use meaningless words. Rather, the learning of real-world languages involves the simultaneous acquisition of grammar rules and word meanings, or semantics. In this study, we have added meaningful stimuli to the standard AGL paradigm in order to assess semantic learning alongside grammar learning. Over the course of 4 testing runs, English-fluent participants learned an artificial language with one of three word orders: an English word order (subject-verb-object, SVO); a widely used, non-English word order (subject-object-verb, SOV); and a rare non-English word order (object-subject-verb, OSV). Many of the participants were multilingual and familiar with at least one SOV language as well, which let us examine how varying word order familiarity affected the learning of an artificial language. No effect was observed, a finding that may be the result of a small sample size. Although the OSV group had the most difficulty learning the language, participants were able to learn the artificial language to a high level of proficiency in all word order conditions. They acquired the grammar and semantics simultaneously, but grammar learning was slower than semantic learning. Interestingly, participants' ability to read and comprehend the language differed from their ability to produce it. The inclusion of semantics presents many important possibilities for future research on language learning and use.

Artificial Language Learning:
The Concurrent Acquisition of Word Order and Semantics

By

Mikael Xie

Ben Wilson

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Psychology

2022

Acknowledgements

I would like to thank the members of Wilson Lab,
who were kind enough to answer all of my dumb questions.

Table of Contents

Introduction.....	1
Pilot Experiment.....	4
Principal Study.....	19
References.....	36

Introduction

Even before we are old enough to understand basic grammar rules, our brains are working below our awareness, memorizing patterns and analyzing the probability that one word or phrase will be followed by another, a process known as statistical learning (Arnon & Snider, 2010; Durrant & Doherty, 2010; Pothos, 2007). Extensive research has demonstrated the high proficiency that human infants have for learning statistical probabilities (Marcus et al., 1999; Saffran & Wilson, 2003; Saffran et al., 1996). This phenomenon is not unique to humans – monkeys (Fitch & Hauser, 2004; Saffran et al., 2008; Wilson et al., 2013) and even certain species of birds are capable of statistical learning (Gentner et al., 2006; Herbranson & Shimp, 2003).

When assessing statistical learning, artificial grammar learning (AGL) paradigms are the most commonly used method. In AGL experiments, subjects are exposed to “sentences” composed of symbols or meaningless nonsense words, which follow particular rules or patterns determined by the grammar. They are then tested on how well they learn the grammatical rules that govern the order of words in these sentences. The convention is to test their new knowledge using a grammaticality judgement task, in which they are presented with sequences and asked to judge whether the sequences are grammatical or ungrammatical. For decades, the AGL paradigm has been used to assess the acquisition of statistical probabilities because it allows for an isolated assessment of syntactic learning while controlling for all other processes – lexical, semantic, and phonological. Artificial grammar learning has also been used to explore the mechanisms underlying conditions such as Parkinson’s disease, aphasia, and dyslexia (Parkinson’s: Smith et al., 2001; aphasia: Hoen et al., 2003; dyslexia: Rüsseler et al., 2006). People who develop these

language disorders exhibit deficits in artificial grammar learning, suggesting that the statistical learning assessed by these paradigms is in an important prerequisite for language.

Despite the extensive literature surrounding artificial grammar learning, AGL is seldom assessed in conjunction with semantic learning, and participants are never asked to generate the sequence themselves. Although there is use in isolating grammar acquisition, it is important to remember real-world languages contain both semantic content and grammar structures. The learning of a language is dependent on an individual's ability to synthesize both semantic and grammatical information (Gibson et al., 2013). Not only is it necessary to understand the syntactic organization of word classes such as nouns, verbs, and adjectives (Smith & Yu, 2008), individual words often have unique grammar rules for how they must fit into a sentence (Goldberg & Suttle, 2010). Furthermore, real-world language learning necessitates the ability to speak or write the language, not just understand it. In this study, we took the traditional artificial grammar learning paradigm one step closer to natural language by evaluating the concurrent acquisition of semantics and a real-world grammar. Although traditional AGL tasks must assess learning by relying on grammaticality judgement tasks, the addition of semantics in our experiment allowed us to instead use a forced choice task, which has been shown to promote active statistical learning (Frinsel et al., 2020). In our case, we used a two-alternative forced choice task (2AFC) in which participants were asked to interpret two written sentences and select the one that best described a visual scene. In addition to the 2AFC task, which assessed participants' ability to interpret and understand the language, we also used a sentence generation task in which participants were asked to produce sequences that described a visual scene. In this way, we were able to assess both comprehension and production abilities.

For the sake of clarity, we will refer to an artificial grammar with semantic content as an artificial language. In the sparse literature that currently examines artificial language learning, the addition of semantics has been found to contribute to high levels of learning of grammars containing hierarchical center embeddings (HCE), center embedded recursion (CER) and crossed dependencies – all of which are grammars commonly known for their difficulty (HCE: Poletiek, 2021; CER: Fedor et al., 2012; Wilson et al., 2020; crossed dependencies: Wilson et al., 2020). Therefore, we predicted that over the course of several testing runs our artificial language would be successfully learned by participants.

In its simplest form, the standard English sentence is composed of a subject and a predicate in the form of a verb (Huddleston et al., 2022). Since we were attempting to create an artificial language as close to natural language as possible but still control for possible confounds, we opted for a simple, real-world grammar that is applied universally: word order – specifically, the word order of a basic transitive sentence containing one subject (S), one object (O), and one verb (V). The vast majority of real-world languages adhere to one of six order variations of subject, object, and verb. Even among languages that can be considered to have freer word orders, there is evidence that one word order predominates (Comrie, 1981).

The use of a real-world grammar results in the possibility that that learning can be biased by previous experience, since participants' native languages have been shown to affect the way they segment artificial languages (Caldwell-Harris et al., 2015; Trecca et al., 2019). To account for this possibility, we chose the following three word orders to use in this study: 'subject-object-verb' (SOV), 'subject-verb-object' (SVO), and 'object-subject-verb' (OSV). SOV and SVO are the most common among natural languages (Greenberg, 1963), and OSV is the least common (Derbyshire & Pullum, 1981). All participants were adults fluent in English, an SVO language,

and some participants were familiar with SOV languages. No participants were familiar with any OSV languages. All participants in the SVO group were predicted to perform well due to their fluency in English, but the performance of participants in the SOV group was expected to vary based on their language background – the higher the familiarity with SOV languages, the easier they would learn the SOV artificial language. Participants in the OSV groups were expected to perform the worst, since none of the participants had familiarity with the OSV word order.

Our purpose in this study was to observe the synchronous learning of grammar and semantics. We predicted that participants would be able to learn both the semantic content and grammar rules of our artificial language over several testing runs. We also hypothesized that if participants did not learn semantics and grammar synchronously, then they would need to learn the meanings of the stimuli before they could organize them into the correct grammar structures. Additionally, if participants' language experiences influence artificial language learning, then with the grammar would facilitate learning. To test participants' ability to apply their knowledge of the artificial language, we included a sentence generation task in addition to the forced choice task. We predicted a strong positive correlation between scores in the forced choice task and scores in the generation task.

After initial data collection, it became clear that participants were reaching ceiling performance too quickly for any meaningful assessment of learning speed or comparison between word order groups. As a result, only 10 participants were tested before the experimental protocol was reworked to increase task difficulty and to decelerate learning (see Principal Study). The data from the initial experimental design was treated as a pilot test preceding the principal study.

Pilot Experiment

Method

Participants

Our research protocol was approved by the Institutional Review Board (IRB). Emory University Psychology students were recruited via Sona-Systems and compensated with course credit. Participants who had any level of fluency in American Sign Language or other ‘object-subject-verb’ (OSV) languages were excluded during data collection, as participants needed to be completely unfamiliar with OSV word order. Before the experiment, participants were asked to provide their age, gender, languages known, and fluency level. We recruited 10 Emory University undergraduates (2 men, 8 women; ages 18-21), representing a population of adults that have high familiarity with a ‘subject-verb-object’ (SVO) language. Participants knew an average of 1.3 SOV languages ($SD = 1.27$). They were also asked to rank their fluency with each language on a scale from 1 to 5 (1 = complete beginner, 5 = native level of fluency), and participants’ average SOV fluency was $M = 1.84$, $SD = 1.63$. Participants were then randomly assigned to the SVO, SOV, or OSV word order group.

Materials

All stimuli were visually presented on a touch-screen computer. The words used in the study were three-letter, consonant-vowel-consonant nonsense words that mimic English morphemes, the smallest meaningful word subdivisions, in order to be pronounceable to participants (e.g., *bif, jat, pob*). To give the words semantic value, the experiment used a vocabulary of nonsense words that were placed into to one of two categories: verbs or nouns. ‘Nouns’ described abstract shapes and ‘verbs’ described interactions between the shapes (e.g., Shape 1 ‘bounces on’ Shape 2; Figure 1 & Table 1). This vocabulary was used to generate ‘sentences’ of three words. All sentences contained one verb, a subject, and an object in one of

three arrangements – SVO, SOV, or OSV word order. For example, the sentence, ‘Shape 1 bounces on Shape 2,’ could be presented as “*Bif pok sut*” (SOV group), “*Bif sut pok*” (SVO group), or “*Pok bif sut*” (OSV group). Each sentence corresponded with a visual ‘scene’ in which two shapes interact: a scene was composed of one moving shape (the ‘subject’) and one motionless shape (the ‘object’); the ‘verb’ corresponded with the path of movement of the ‘subject.’ Although the ‘verbs’ in the figures and tables are represented by arrows (Table 1 & Figure 2), participants were shown genuine movements. There were 24 total scene-sentence variations.


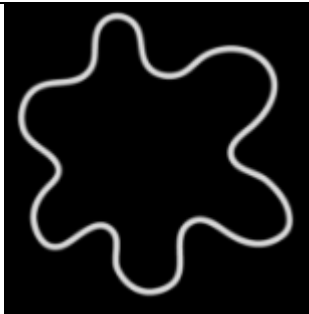



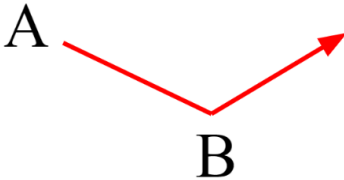
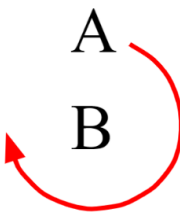
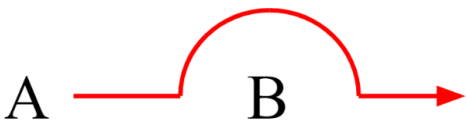
Noun	Nonsense Word
 <p data-bbox="305 667 412 705">Shape 1</p>	<p data-bbox="662 520 727 571"><i>bif</i></p>
 <p data-bbox="305 1035 412 1073">Shape 2</p>	<p data-bbox="651 888 732 938"><i>pok</i></p>
 <p data-bbox="305 1392 412 1430">Shape 3</p>	<p data-bbox="662 1245 721 1295"><i>jat</i></p>
 <p data-bbox="305 1717 412 1755">Shape 4</p>	<p data-bbox="651 1591 732 1642"><i>wum</i></p>

Figure 1. Noun Stimuli and Corresponding Nonsense Words. Shapes 1-3 replicated from Kirby et al., 2015.

Table 1*Verb Stimuli and Corresponding Nonsense Words*

Verb	Action	Nonsense Word
Hit	 <p data-bbox="495 451 1079 483">“A moves towards B until both objects touch”</p>	cax
Bounce On	 <p data-bbox="479 714 1104 787">“A moves down and towards B until both objects touch, then up and away from B”</p>	sut
Circle	 <p data-bbox="584 1050 990 1081">“A moves clockwise around B”</p>	lis
Jump Over	 <p data-bbox="462 1260 1120 1333">“A moves towards B, in a semicircle around B, and then away from B”</p>	yoz

Note. Movement is denoted by the red arrow with a short description in quotations below. Shapes are represented by placeholders “A” and “B”.

Procedure

The Pilot Experiment was composed of three parts: an exposure phase, followed by a two-alternative forced choice (2AFC) task, and then a generation task. Four total runs of exposure, choice, and generation were conducted. All stimuli were presented visually on a computer screen with no auditory components.

Exposure Phase.

First, participants were familiarized with the artificial language. To start, participants were seated in front of a touch screen computer with the following instructions displayed: "Please carefully watch the following scenes and sentences. Try your best to learn the language and its rules. Tap the screen to start." Then, in each run of exposure, participants were serially shown the same 24 scenes depicting two shapes interacting with each other, as well as the corresponding sentence presented visually below the scene (see Figure 2A). Scene order was randomized for every run of Exposure. After the 'subject' completed its movement, each scene remained on screen for one second before moving on to the next scene. The 24 sentences were presented so that each of the four nouns was used as a subject six times and an object six times; each of the four verbs was used six times as well. No sentences were repeated, and no words were repeated within sentences. No participant input was required for this phase.

Two-Alternative Forced Choice (2AFC) Task.

Next, participants were assessed on their knowledge of the artificial language. The task began with an instruction screen: "Please look at the following scenes. Select the sentence that correctly describes the scene. Tap the screen to start." Participants were then presented with a novel visual scene and two sentences, and they attempted to select the correct sentence out of the two options by tapping the screen with their hand (see Figure 2B). No time limit was enforced, but participants could not advance to the next scene until they had selected one of the two

sentences. The incorrect sentence differed from the correct sentence by one of the following features: subject, verb, object, or word order. For example, “*wum bif lis*,” as an SOV sentence, could have foils of “*jat bif lis*” (incorrect subject), “*wum bif sut*” (incorrect verb), “*wum pok lis*” (incorrect object), “*wum lis bif*” (incorrect word order – verb location changed), or “*bif wum lis*” (incorrect word order – nouns swapped). Incorrect subject, object, and verb trials were classified as tests of vocabulary learning, and incorrect order trials were tests of grammar learning. This allowed us to test for differences between vocabulary and grammar learning. Participants were given immediate feedback following their responses in the form of a blank red or green screen, since both positive and negative feedback are beneficial for artificial language learning (Frinsel et al., 2020). Participants were expected to understand that the red screen indicated they were incorrect, and the green screen indicated they were correct; no other information or feedback was given before moving onto the next scene. Each run of the 2AFC task contained 24 scene variations that were not shown in the exposure phase, presented in a random order. The sentences in the 2AFC phase were developed with the same rules used to construct sentences for the exposure phase (see Exposure Phase).

Generation Task.

The final phase gave a more detailed understanding of what participants learned during the previous phases and their ability to apply that knowledge by assessing their ability to produce language. Each generation phase began with the same instruction screen: "Please look at the following scenes. Generate a sentence that correctly describes the scene. Tap the screen to start." Participants were presented with the entire nonsense vocabulary as well as a visual scene, and were required to select, in order, words to describe the scene (see Figure 2C). The organization of the nonsense vocabulary on the screen was randomized for each participant prior to Run 1 to

avoid any position biases. To prevent confusion, the word positions were maintained in the original randomized positions for the remainder of the participant's runs. As they selected each word, the chosen words would appear above the scene in the order they were selected (Figure 2C); participants were not able to clear or edit any words once they were selected. Feedback was given after three words were chosen, since participants were exclusively exposed to three-word sentences. Participants were shown a blank green screen if they selected all three correct words in the correct order, and they were shown a blank red screen if any of the words were incorrect or placed in an incorrect position. Once again, no time limit was enforced, but participants could not advance to the next scene until they had completed the sentence. The same 24 scene variations as the 2AFC task were shown in a randomized order.

Scoring and Data Analysis.

Data analysis was performed in SPSS (IBM SPSS Statistics, version 27). We calculated the proportion of correct answers of each run of the 2AFC task, allowing for assessment of how performance changed as the runs progressed. For the generation phase, the performance of each trial was calculated in several ways. Generation 'proportion performance' was calculated as the proportion of the response that contained the correct word in the correct location. For example, an answer of "*lis jat pok*" would receive a score of 0.33 if the correct answer was "*wum lis pok*." This method of scoring was used to represent general performance. 'Vocabulary performance' was scored according to the proportion of correct words the participants chose, regardless of order in which the words were chosen. For example, a response of "*pok cax bif*" would receive a full score of 1 if the correct answer was "*bif pok cax*." 'Grammar performance' was scored according to the proportion of word types (nouns and verbs) placed in the correct position. Note that each participant had one of the following word type organizations: noun – noun – verb (SOV

and OSV) or noun – verb – noun (SVO). A response of “*jat wum yoz*” (noun-noun-verb) would receive a full score of 1 if the correct answer was “*bif pok cax*” (noun-noun-verb), as the grammatical categories were correct, even though the specific words were incorrect.

Using one-sample t-tests with Bonferroni corrections, 2AFC performance overall and by run were compared to chance levels (0.50). Generation performance overall and by run were also compared to chance (0.12) via one-sample t-tests with Bonferroni corrections. Chance levels were calculated by running 1,000,000 simulations of randomly generated responses in each task and averaging the scores. Since we expected participants to learn the artificial language, we predicted that overall performance in both tasks would be significantly above chance, and performance would increase across testing runs.

The data from the 2AFC and generation tasks were analyzed with two-factor ANOVAs: factors were defined as Run Number and Word Order, with the dependent variable of performance. We predicted a main effect for Run Number, corresponding to an increase in scores over time in both 2AFC and generation tasks. We also predicted a main effect of Word Order, with performance being the highest for the SVO, lower for SOV, and lowest for OSV. A Pearson’s correlation test was performed to evaluate how well 2AFC scores predicted generation scores, and a strong positive correlation was predicted.

Another set of two-factor ANOVAs was performed to compare semantic and grammar learning. 2AFC performance data were run with the following factors: Run Number and Foil Type. We predicted a main effect of Foil Type, showing a difference in performance between questions with vocabulary foils versus those with grammar foils. An interaction effect was expected as well, aligning with our hypothesis that vocabulary is learned faster than grammar. Participants were expected to need some semantic knowledge to figure out how the shapes and

movements in the scene corresponded with the arrangement of words in the sentence below. The generation data were run with the factors of Run Number and Scoring Type. Scoring Type was composed of two levels: Vocabulary Trials and Grammar Trials. Since chance levels in Vocabulary Trials (0.38) were different from chance levels in Grammar Trials (0.50), any main effect of Scoring Type would be disregarded if it could be attributed to this difference (i.e., if Grammar Trial performance was higher than Vocabulary Trial performance). More importantly, we predicted an interaction effect, meaning the rate of learning vocabulary learning would differ from the rate of grammar learning.

To test the effect of language experience on the learning of our artificial language, we used a Pearson's Correlation test to assess how effectively SOV proficiency predicted performance in the SOV group. SOV proficiency for each participant was calculated by adding together the self-reported levels of fluency for each SOV language they knew. For example, a participant who knew two SOV languages and recorded their fluency levels as 2 and 4 (out of a maximum fluency level of 5) would have a SOV proficiency of 6. We predicted a positive correlation between SOV proficiency and performance within the SOV group.

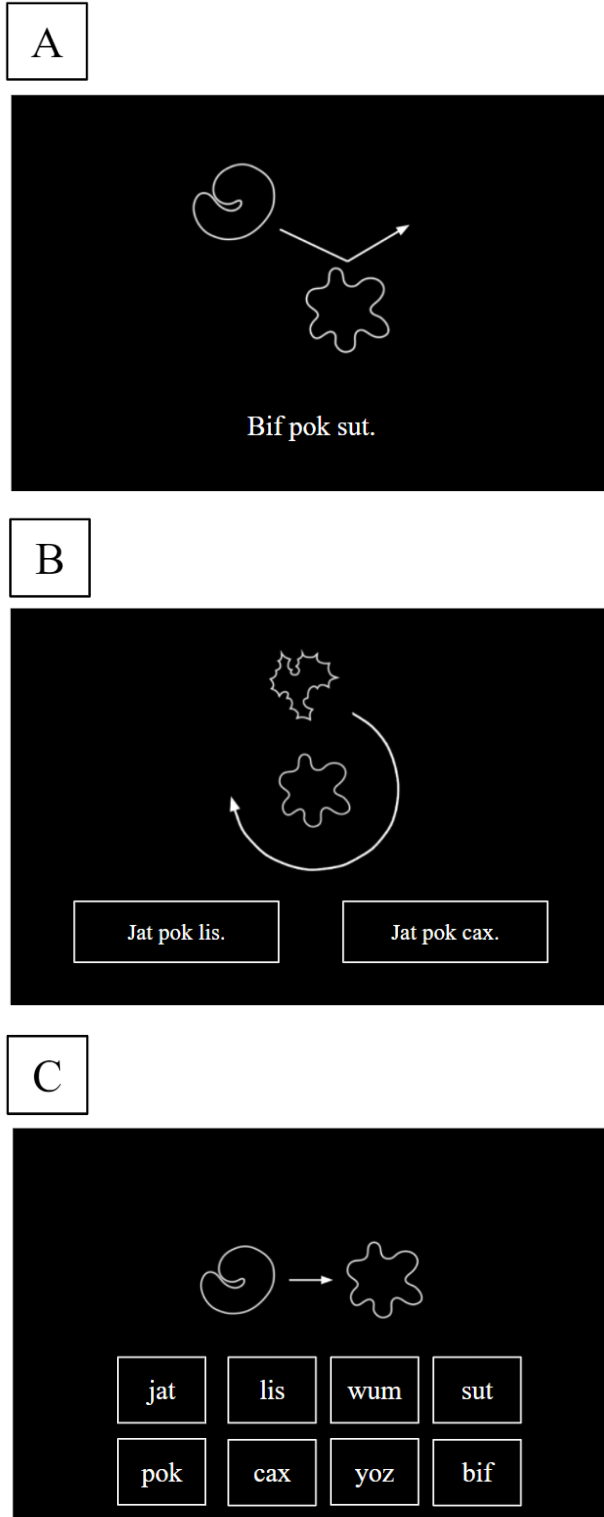


Figure 2. Example of Participant View During Exposure (A), 2AFC (B), and Generation (C) Phases. Participant selection options are denoted by white boxes. Movement is represented by arrows, but participants viewed actual movements.

Results

With one exception, all participants learned the artificial language with ease, reaching ceiling level by Run 2 (Figure 3A). Initial analyses revealed that performance in the pilot experiment reached ceiling levels too quickly for us to measure learning over time, so data collection was halted after testing 10 participants. Nevertheless, several planned analyses were conducted. Starting with a one-sample t-test, we found that overall 2AFC performance was high above chance ($M = 0.91$, $SD = 0.15$, $t_9 = 8.60$, $p < .001$, $d = 2.72$). A two-way ANOVA showed no main effect of Run ($F_{1.15,8.06} = 4.06$, $p = .075$, $\eta_p^2 = .367$), no main effect of Group ($F_{2,7} = 0.55$, $p = .598$, $\eta_p^2 = .137$; Figure 3B), and no interaction effect ($F_{2.30,8.06} = 0.39$, $p = .088$, $\eta_p^2 = .100$). The Generation task showed similar results – overall performance was far above chance ($M = 0.86$, $SD = 0.27$, $t_9 = 9.23$, $p < .001$, $d = 2.92$). No main effects were found for Run ($F_{1.03,7.23} = 3.77$, $p = .091$, $\eta_p^2 = .350$; Figure 4A) or Group ($F_{2,7} = 1.14$, $p = .372$, $\eta_p^2 = .246$; Figure 4B), and no interaction effect was found ($F_{2.07,7.23} = 0.49$, $p = .640$, $\eta_p^2 = .122$). Average performance in both tasks was at or near ceiling level across all runs and all groups, meaning the task was learned too quickly for any meaningful assessment of learning. Rather than continuing to collect more data, we revised the experiment by removing the Exposure phase, which proved to be effective at depressing learning speed (see Principal Study).

Discussion

The results from this pilot experiment showed the artificial language was easily learned by Run 2, as shown in both the 2AFC and generation phases. Possible explanations for this high level of performance include the lengthy Exposure Phase at the start of every run, which gave copious time for participants to observe the language rules, and the inclusion of both negative and positive feedback, which allowed for active learning during the 2AFC task (Frinsel et al.,

2020). Due to the high level at which the participants were performing, we decided to remove the Exposure Phase to make the task more difficult and potentially produce more gradual learning.

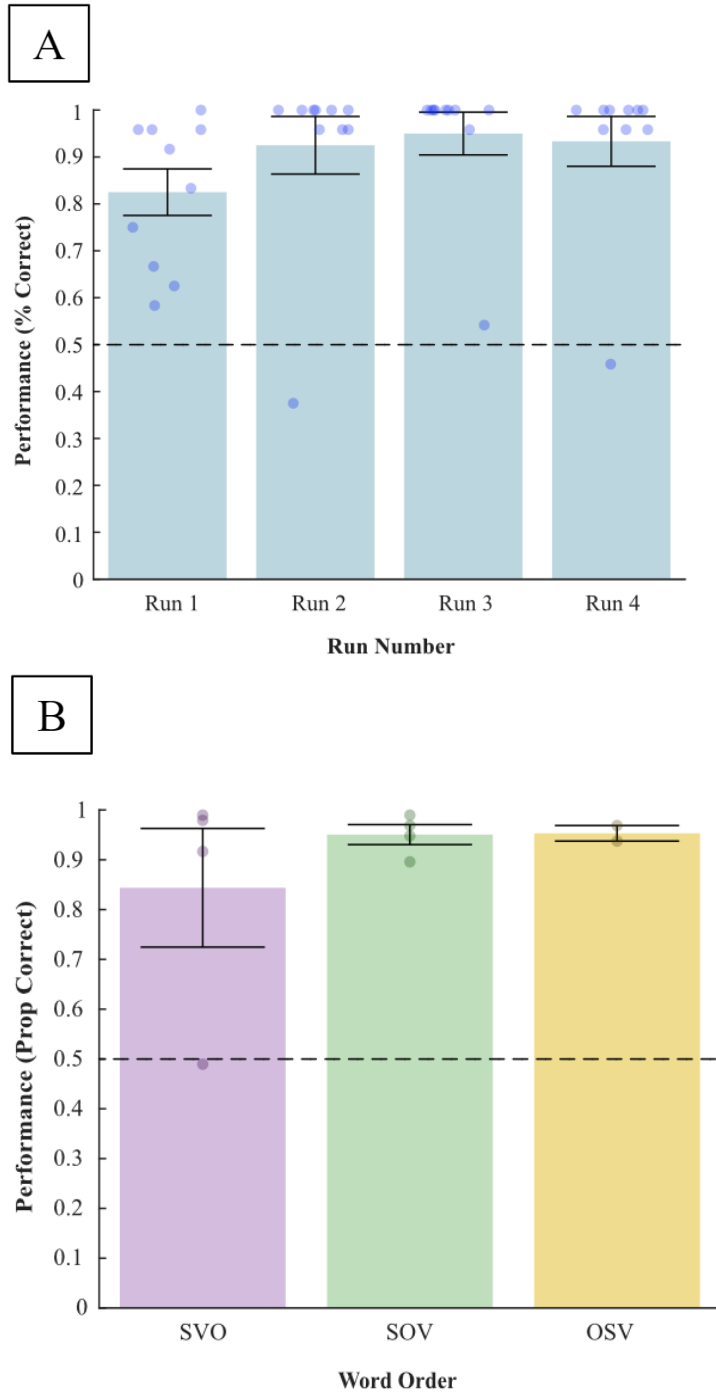


Figure 3. 2AFC Performance Across Runs (A) and Word Order Groups (B) in the Pilot Experiment. Chance levels (0.50) are denoted by dashed lines. Performance reached ceiling levels by Run 2 (A) and did not differ significantly across groups (B).

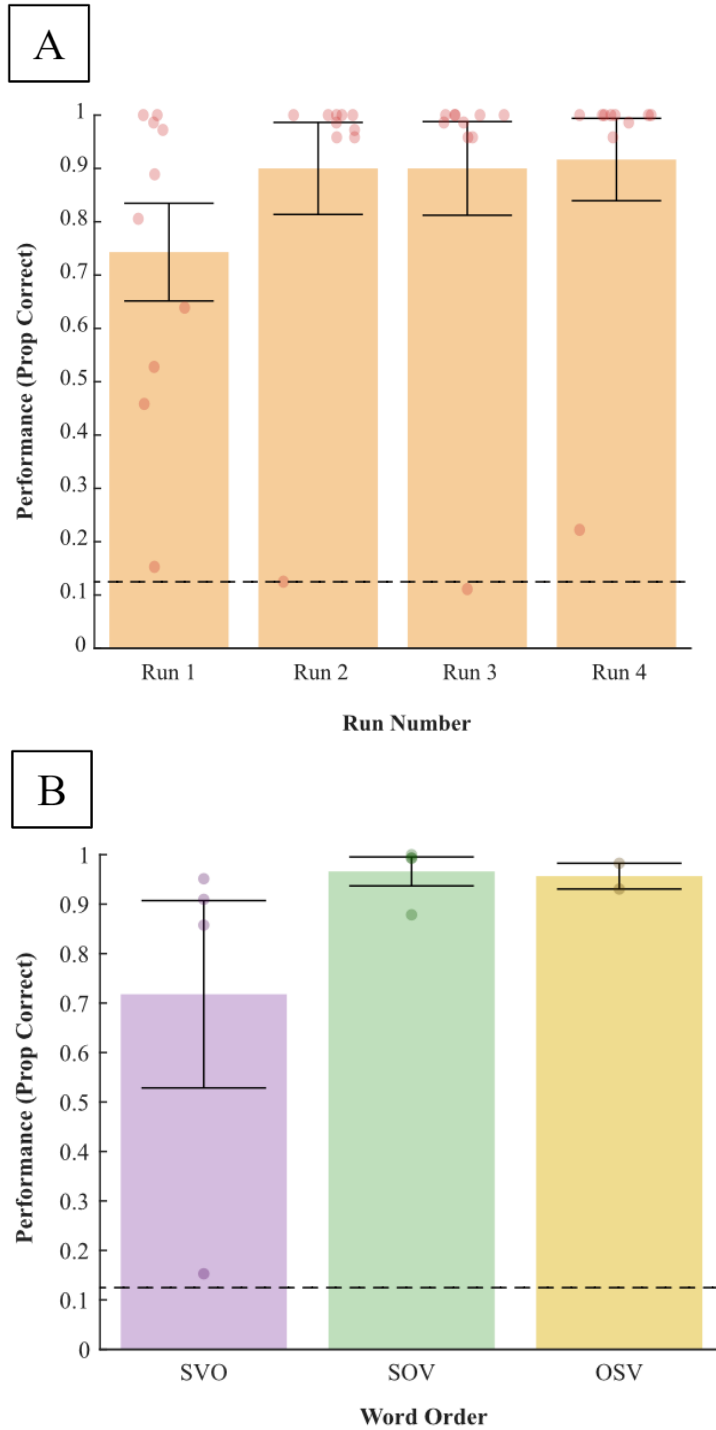


Figure 4. Generation Performance Across Runs (A) and Word Order Groups (B) in the Pilot Experiment. Chance levels (0.12) are denoted by dashed lines. Like 2AFC (see Figure 3), generation performance reached ceiling level by Run 2 (A) and did not vary significantly across groups (B).

Principal Study

Method

Since some participants scored at or near 100% in Run 1 of the 2AFC Task in the Pilot Experiment (see Figure 3), it was clear that the extensive exposure phase was a significant source of learning. In an attempt to slow down language learning, we removed the exposure phase. As a result, participants would not receive any information about the artificial language before they were tested on language knowledge. Starting with the forced choice task, they would be required to learn the artificial language via trial and error, a method that Frinsel et al. (2020) argued is more effective for the assessment of language learning because it provides more insight into the trajectory of learning than a passive exposure phase. The forced choice task not only allows us to observe gradual acquisition of the language via the participants' answer choices, it promotes active learning of the language via immediate feedback. Typically, second languages are not learned through passive exposure but require active learning, so this alteration in our protocol could shift the paradigm closer to real-world additional language acquisition. Moreover, the exposure phase at the start of every run served as a reminder of what a correct, grammatical sentence looks like when matched to its scene pair, and without this reminder, participants were expected to have a more difficult time learning the artificial language.

Participants

34 Emory University undergraduates (ages 18-21; 7 men, 28 women) were recruited via Sona-Systems and compensated with course credit. Participants with any level of fluency in American Sign Language or other 'object-subject-verb' (OSV) languages were not eligible for the study. Participants knew an average of 0.97 SOV languages ($SD = 0.92$). Participants' average SOV fluency (see Pilot Experiment, participants) was $M = 2.54$, $SD = 2.74$. Twelve were

given the SOV language, 11 were given the SVO language, and 11 were given the OSV language.

Materials

The same materials from the Pilot Experiment were used (see Figure 1 & Table 1).

Procedure

The principal study was composed of only the forced choice and generation tasks; four total runs were conducted. Prior to the first 2AFC phase, participants were asked to provide their age, gender, native language, languages known, and fluency level. Participants were then randomly assigned to one of the SVO, SOV, or OSV word order groups. The forced choice and generation phases remained unchanged from the Pilot Experiment, but this time participants were verbally informed to try to learn the language via trial and error.

Scoring and Data Analysis.

Performance scoring and data analysis were conducted as described in the Pilot Experiment.

Results

As we predicted, participants on average were able to successfully learn each of the artificial languages. Four one-sample t-tests were performed with Bonferroni corrections, comparing average 2AFC performance in each run to chance (0.50). Performance in the 2AFC task started significantly above chance in Run 1 ($M = 0.60$, $SD = 0.14$, $t_{33} = 4.27$, $p < .001$, $d = .73$). Run 2 ($M = 0.71$, $SD = 0.22$), Run 3 ($M = 0.81$, $SD = 0.21$), and Run 4 ($M = 0.85$, $SD = 0.22$) were all significantly above chance as well (Run 2: $t_{33} = 5.55$, $p < .001$, $d = .95$; Run 3: $t_{33} = 8.67$, $p < .001$, $d = 1.49$; Run 4: $t_{33} = 9.037$, $p < .001$, $d = 1.55$). In all four runs, 2AFC performance was significantly higher than chance, suggesting learning occurred (Figure 5A).

Three additional one-sample t-tests with Bonferroni corrections were run for 2AFC performance in each word order group (Figure 5B). The SVO group ($M = 0.81$, $SD = 0.13$, $t_{10} = 7.44$, $p < 0.001$, $d = 2.24$), the SOV group ($M = 0.79$, $SD = 0.14$, $t_{11} = 6.70$, $p < 0.001$, $d = 1.94$), and the OSV group ($M = 0.63$, $SD = 0.16$, $t_{10} = 2.52$, $p = .030$, $d = .768$) all performed above chance in the 2AFC Task.

A two-way ANOVA (factors: Run and Word Order) revealed a strong main effect of Run ($F_{2,35,72.82} = 26.54$, $p < .001$, $\eta_p^2 = .461$; Figure 5A) and a strong main effect of Word Order ($F_{2,31} = 4.47$, $p = .02$, $\eta_p^2 = .224$; Figure 5B). No significant interaction effect was found between Word Order and Run ($F_{4,70,72.82} = 2.03$, $p = .088$, $\eta_p^2 = .116$; Figure 6). Post hoc tests were conducted with the Bonferroni correction, and it was found that 2AFC performance was significantly different between Runs 1 and 2 ($p = .016$, 95% C.I. = [-0.20, -0.02]), as well as Runs 2 and 3 ($p = 0.012$, 95% C.I. = [-0.19, -0.02]). There was no statistically significant difference in performance between Runs 3 and 4 ($p = 0.468$), since participants were nearing ceiling level (Figure 5A). 2AFC performance was significantly higher in the SVO group than in the OSV group ($p = .032$, 95% C.I. = [0.01, 0.34]). Interestingly, although the SVO and SOV groups had nearly equivalent scores ($p = 1.000$, with Bonferroni correction), the difference between the SOV and OSV groups was not quite significant ($p = 0.058$; Figure 5B). This result suggests that if more participants had been tested, it is possible that a difference between the SOV and OSV groups would have been detected.

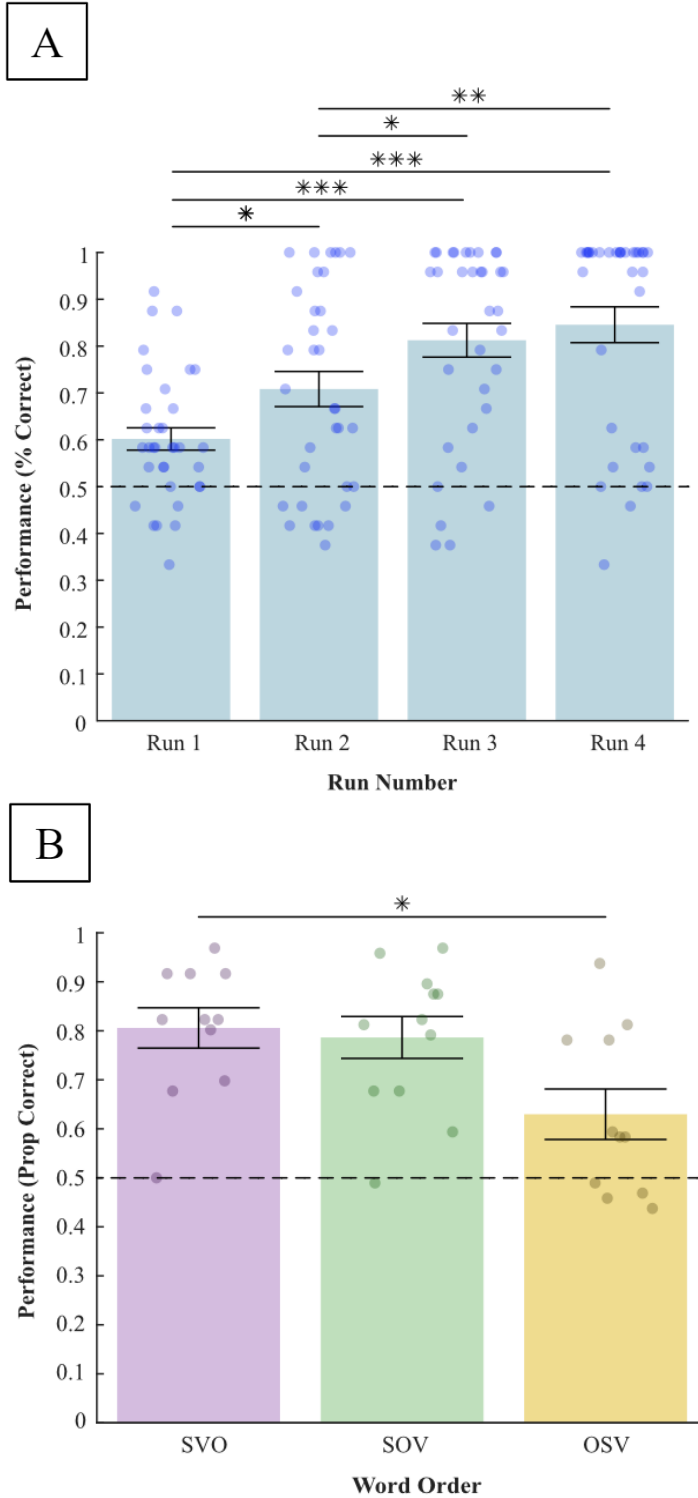


Figure 5. 2AFC Performance Across Runs (A) and Word Order Groups (B) in the Principal Study. Chance levels (0.50) are denoted by dashed lines. (A) Performance in all runs were significantly above chance, and performance increased with each run. (B) All groups performed above chance, but there was a significant difference between SVO and OSV groups.

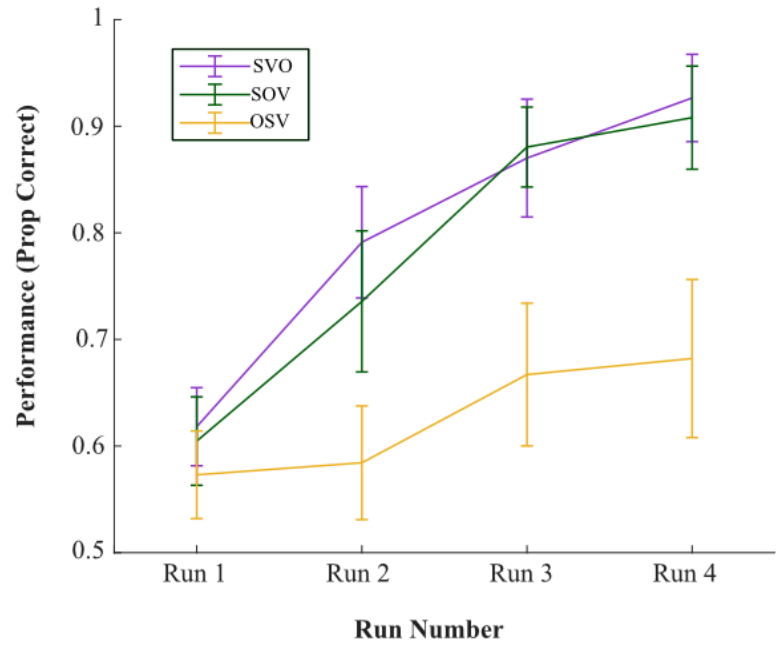


Figure 6. Interaction between Word Order and Run in the 2AFC Task (Principal Study). The x-axis is positioned at chance (0.50). There were no significant differences in the rate of learning across word order groups.

In alignment with the 2AFC data, we also predicted that participants would perform well in the generation task, and that performance would increase over time. Four one-sample t-tests with Bonferroni corrections were performed, comparing average Generation performance across runs to chance (0.12, see Pilot Experiment, Scoring and Data Analysis). Run 1 ($M = 0.36$, $SD = 0.33$) was significantly higher than chance, ($t_{33} = 4.04$, $p < .001$, $d = .707$). Average scores in Run 2 ($M = 0.54$, $SD = 0.38$), Run 3 ($M = 0.71$, $SD = 0.37$), and Run 4 ($M = 0.74$, $SD = 0.37$) were all above chance as well (Run 2: $t_{33} = 6.27$, $p < .001$, $d = 1.09$; Run 3: $t_{33} = 9.21$, $p < .001$, $d = 1.59$; Run 4: $t_{33} = 9.70$, $p < .001$, $d = 1.68$). Three more one-sample t-tests with Bonferroni corrections were then performed to analyze Generation performance across word order groups (Figure 7B). All groups performed significantly above chance. Once again, the SVO group ($M = 0.81$, $SD = 0.14$, $t_{10} = 16.56$, $p < .001$, $d = 5.03$); the SOV group ($M = 0.79$, $SD = 0.15$, $t_{11} = 15.46$, $p < .001$, $d = 4.50$); and the OSV group ($M = 0.63$, $SD = 0.17$, $t_{10} = 9.81$, $p < .001$, $d = 2.99$) all scored significantly above chance.

A two-way ANOVA showed a main effect of Run ($F_{1,70,52.70} = 14.27$, $p < .001$, $\eta_p^2 = .443$), suggesting that performance increased across runs (Figure 7A), and a main effect of Word Order ($F_{2,31} = 4.11$, $p = .026$, $\eta_p^2 = .209$), suggesting that there was a significant difference in learning across groups (Figure 7B). Again, no interaction effect was found between Word Order and Run ($F_{3,40,72.82} = 2.07$, $p = .108$, $\eta_p^2 = .067$; Figure 8). Post-hoc tests with Bonferroni corrections found that there was a significant difference between Runs 1 and 2 ($p = 0.042$, 95% C.I. = [-0.15, -0.002]; Figure 7A). The difference between Runs 2 and 3 was not quite significant ($p = 0.051$, 95% C.I. = [-0.15, 0.00]), but Runs 2 and 4 were significantly different ($p = 0.035$, 95% C.I. = [-0.16, -0.004]). Participants scored similarly in Runs 3 and 4 ($p = 1.000$), which suggests that performance in Run 3 and 4 was near ceiling. Generation performance across Word

Order (Figure 7B) aligned with the 2AFC data. The SVO group had scores similar to the SOV group ($p = 1.000$). Although the OSV group scored significantly lower than the SVO group ($p = 0.036$, 95% C.I. = [0.02, 0.65]), the OSV group did not score differently from the SOV group ($p = .092$).

Overall 2AFC performance ($M = 0.74$, $SD = 0.17$) was higher and less variable than Generation performance ($M = 0.59$, $SD = 0.32$). A likely explanation for this result was that the 2AFC task was inevitably easier, requiring participants only to select between one of two options (chance level of 0.5). By contrast, generating a correct sequence by random chance was far less likely (chance level of 0.12). A Pearson's correlation test revealed a strong positive correlation between 2AFC and generation performance, ($r(32) = 0.96$, $p < .001$, $R^2 = 0.93$), indicating that 2AFC performance was a highly effective predictor of generation performance (Figure 9).

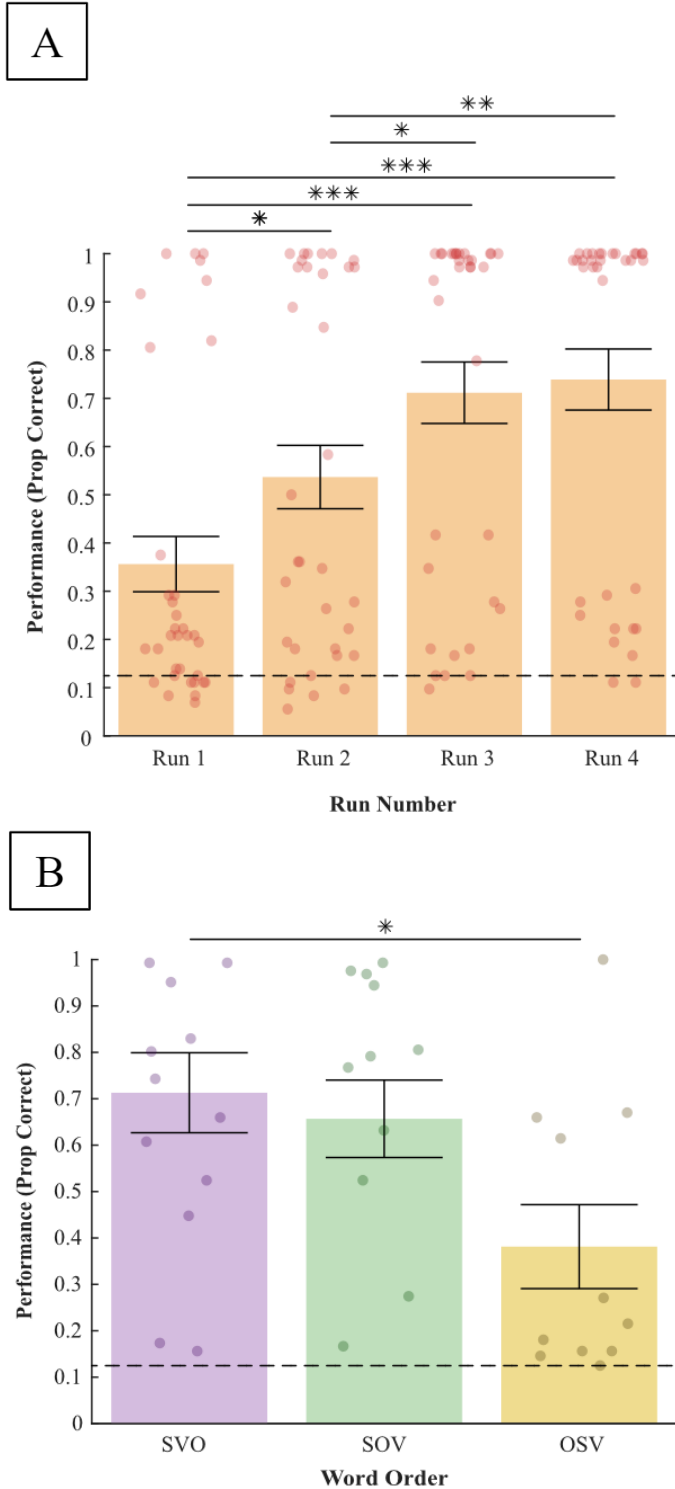


Figure 7. Generation Performance Across Runs (A) and Word Order Groups (B) in the Principal Study. Chance levels (0.12) are denoted by dashed lines. (A) In alignment with 2AFC data, performance in all runs were significantly above chance, and performance increased with each

run. (B) All groups performed above chance and there was a significant difference in performance between SVO and OSV groups.

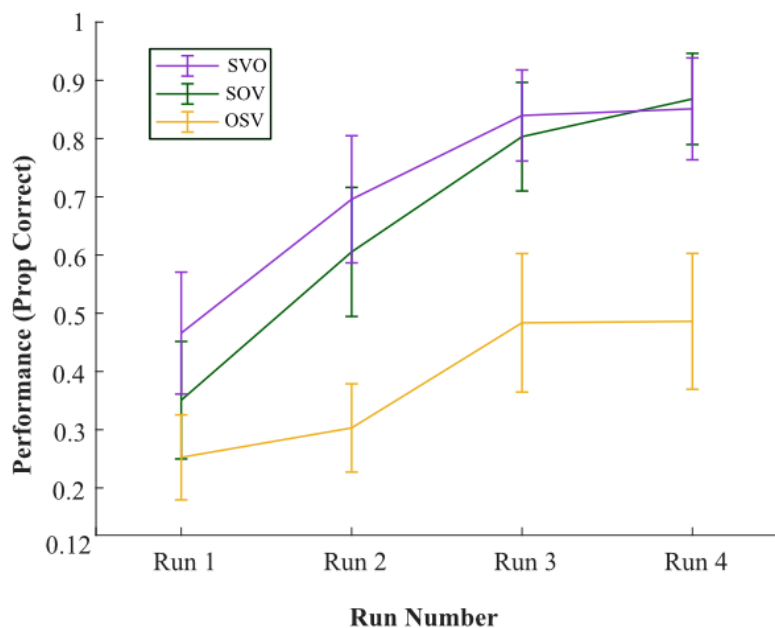


Figure 8. Interaction between Word Order and Run in Generation Task (Principal Study). The x-axis is positioned at chance (0.12). As was the case for 2AFC (see Figure 6), there were no significant differences in the rate of learning across word order groups.

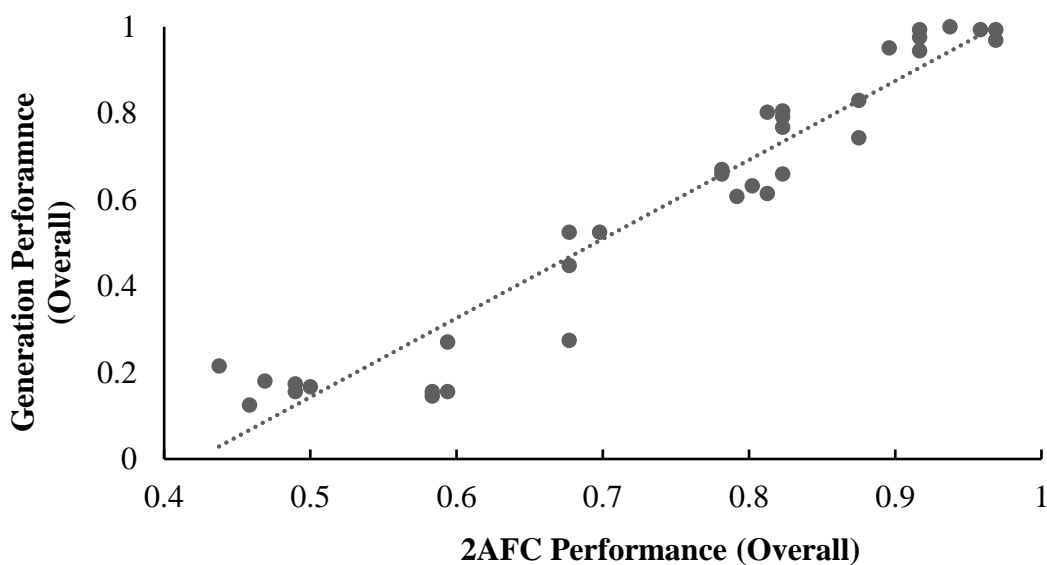


Figure 9. Generation Performance as a Function of 2AFC Performance in the Principal Study. 2AFC performance was strongly correlated with generation performance ($r(32) = 0.96$, $p < .001$, $R^2 = 0.93$).

To analyze the relationship between semantic and grammar learning, a two-way ANOVA was run on the 2AFC performance data with the factors of Run Number and Foil Type, which was composed of Grammar Foils and Vocabulary Foils (Figure 10). We found a main effect of Foil Type ($F_{1,31} = 5.07, p = .032, \eta_p^2 = .140$) and an interaction effect between Run Number and Foil Type ($F_{2,48,76.84} = 5.82, p = .002, \eta_p^2 = .158$), suggesting different rates of learning. Average performance with Grammar Foils was significantly higher than performance on Vocabulary Foils ($p = 0.032, 95\% \text{ C.I.} = [0.003, 0.06]$). This result could be attributed to grammar foil performance starting at a higher level than vocabulary foil performance. Similarly, a two-way ANOVA was run on the Generation performance data (factors: Run Number and Scoring Type, which was composed of Grammar Trials and Vocabulary Trials). A moderate main effect was found for Scoring Type ($F_{1,31} = 4.31, p = .046, \eta_p^2 = .122$), suggesting average vocabulary performance was significantly higher than grammar performance (Figure 11). This finding opposes the 2AFC results, signalling a possible difference between comprehension and generation capabilities. Like the 2AFC task, we found a significant interaction between Run Number and Scoring Type ($F_{1,85,57.46} = 13.11, p < .001, \eta_p^2 = .297$), indicating a difference in learning speed between semantics and grammar.

For our final analysis, we ran a Pearson's correlation test to assess how SOV proficiency related to 2AFC and generation performance within the SOV group (Figure 12). Since all participants were fluent in English, an SVO language, and no participants had any familiarity with an OSV language, no correlation test was necessary for the other two groups. No correlation was found between SOV proficiency and 2AFC performance ($r(10) = 0.09, p = .775, R^2 = 0.009$), nor between SOV proficiency and generation performance ($r(10) = 0.05, p = .888, R^2 = 0.002$). This result may have been a consequence of the low sample size ($N = 12$).

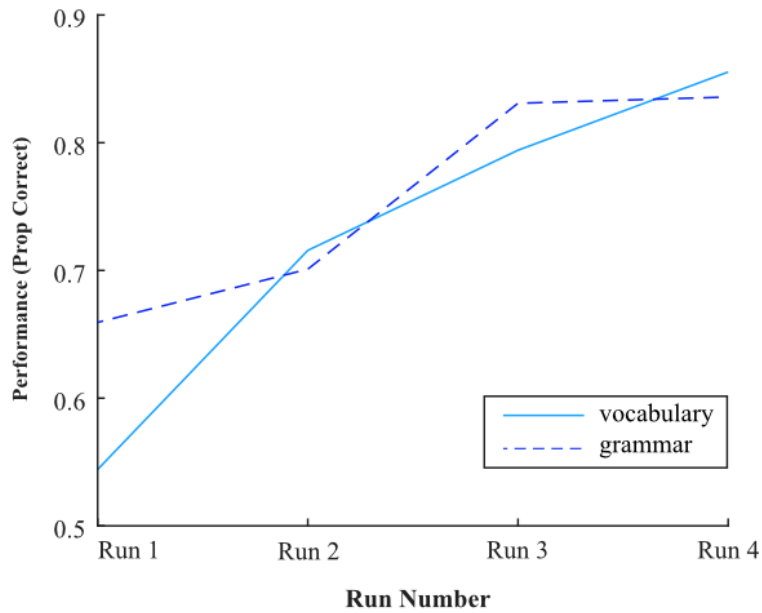


Figure 10. Learning Speed of Vocabulary versus Grammar Structure in 2AFC Task (Principal Study). The x-axis is positioned at chance (0.50). Overall grammar performance was significantly higher than overall vocabulary performance, a finding that was likely attributable to the difference between vocabulary and grammar performance in Run 1. Vocabulary learning was significantly faster than grammar learning.

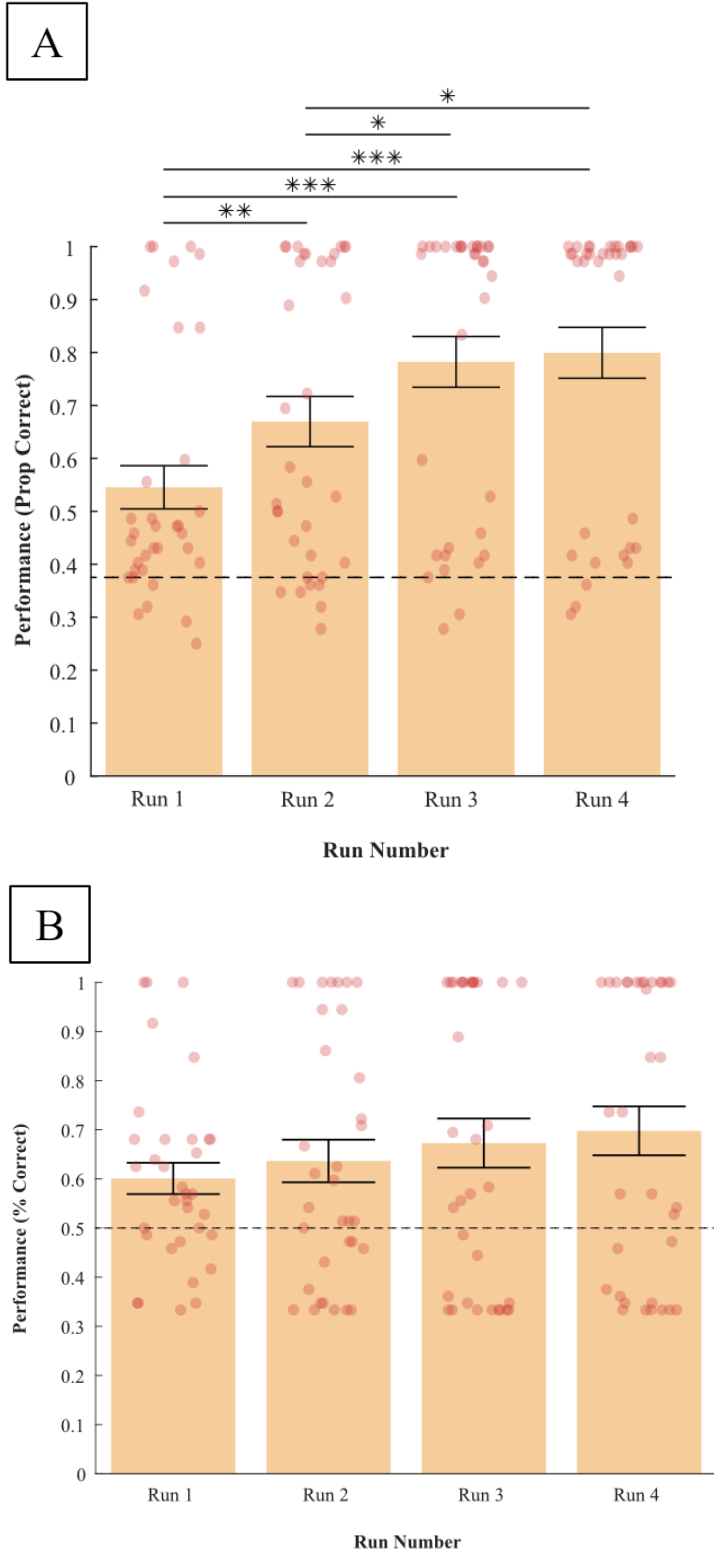


Figure 11. Learning Speed of Vocabulary (A) and Grammar Structure (B) in Generation Phase (Principal Study). Chance levels (A: 0.38; B: 0.50) are denoted by dashed lines. Vocabulary learning (A) was significantly faster than grammar learning (B). Average grammar performance

(B) was significantly lower than average vocabulary performance (A), despite having higher chance levels.

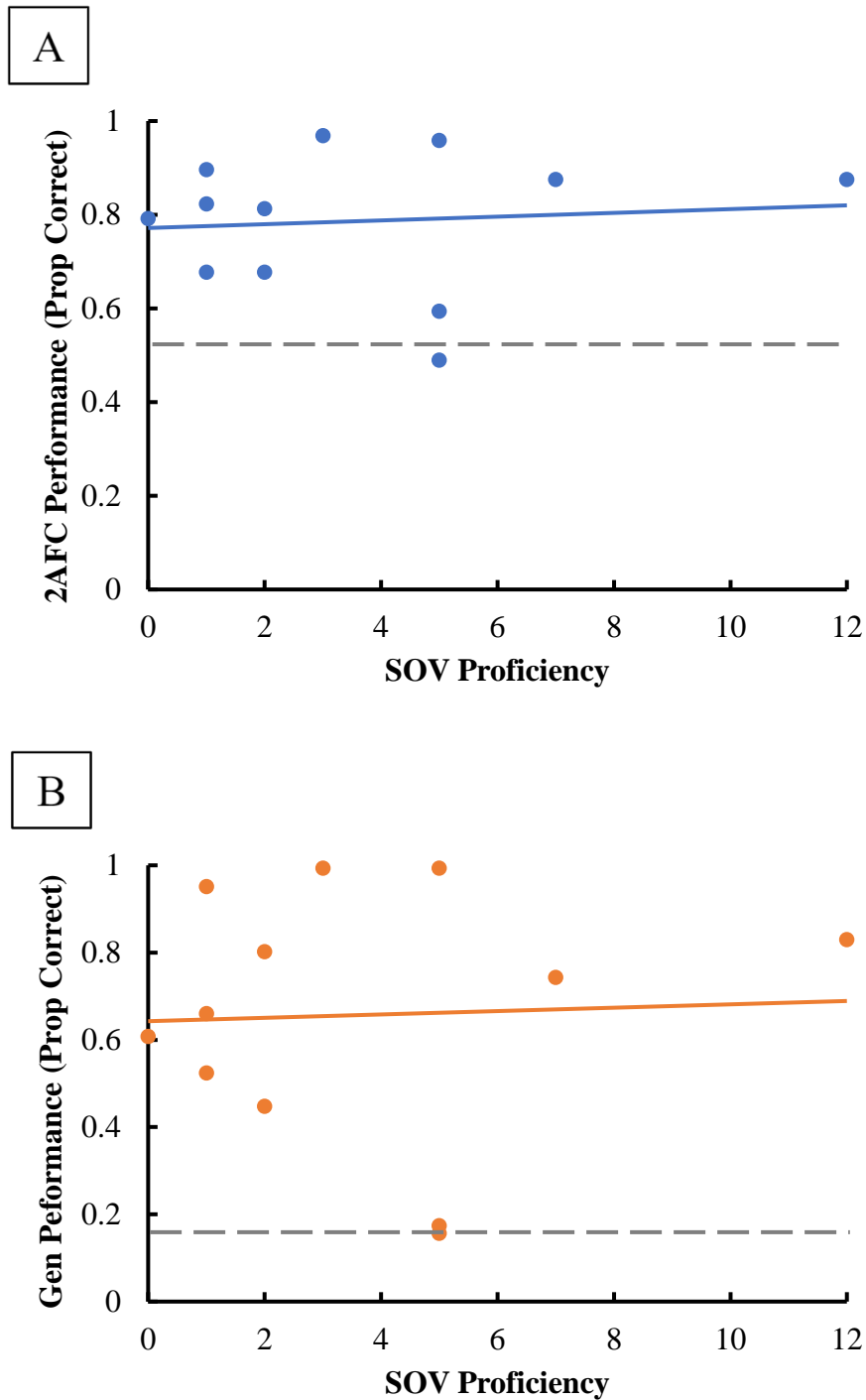


Figure 12. SOV Proficiency as a Predictor of SOV Performance in 2AFC (A) and Generation (B) Tasks (Principal Study). Chance levels (A: 0.50; B: 0.12) are denoted by dashed lines. No correlation between SOV proficiency and performance was found in either task.

Discussion

The addition of semantics into the traditional artificial grammar learning paradigm is a relatively new concept, and to our knowledge, no other study has assessed the learning of an artificial language with this level of similarity to real-world language and tested participants' ability to construct their own sentences. Poletiek et al. (2021) and Wilson et al. (2020) used a vocabulary of nouns and adjectives, which are important components of a sentence but their vocabulary, lacking verbs, does not meet the requirements of a complete sentence (Rodney & Geoffrey, 2005). Frinsel et al. (2020) used a vocabulary of nouns and verbs, but every nonsense word in their vocabulary corresponded with a physical shape, including the verbs (verb words were matched to different arrow shapes). Adult participants certainly would have no trouble associating these shapes with actions; however, verbs in real-world languages are relational terms and cannot be mapped to concrete objects the same way that nouns can, and this contrast has a clear effect on the ways in which verbs are learned differently from nouns (Gentner, 2006). We attempted to address this in our study by presenting the verbs as genuine movements rather than representations of movement. Öttl et al. (2017) did use a vocabulary of nouns (matched to images) and verbs (matched to movements on the screen); however, prior to the grammar learning task, their participants were taught the entire vocabulary of nouns and verbs until they reached 100% proficiency. The researchers did not attempt to assess semantic learning and instead focused on how grammar learning is affected by semantic knowledge; they found no effect.

In our experiment, we assessed how participants learned both the semantic content and grammar rules of the artificial language. The data show that participants were capable of learning the artificial language to a high level of proficiency, nearing ceiling by Run 3. Participants scores

in the 2AFC task – a test of comprehension in which they needed only to pick the correct sentence out of two options– were highly predictive ($r = 0.96$) of their scores in the Generation Task – a test of production in which they needed to generate the sentences on their own. In line with our predictions, grammar rules were acquired at slower rate than the semantic content. Average 2AFC performance on grammar questions was higher than on semantic questions, but average Generation performance on semantic questions was higher. This result may speak to a difference between the ability to recognize a grammatical sentence and the ability to reproduce the grammar. The differences in learning rates and performance between grammar and semantics suggest that grammatical and semantic information can be acquired synchronously, but possibly through two different mechanisms of learning, as proposed by Öttl et al. (2017) and Peña et al. (2002).

In addition to performance across time, we also looked at performance across different word orders. Participants who learned the SVO and SOV languages scored highly and there was no significant difference between the two groups, while the OSV word order was far more difficult for participants to learn than the SVO word order. The difference between OSV performance and SOV performance was not significant; however, with a larger sample size, OSV performance likely would have been found to be significantly lower than both SVO and SOV groups. The difficulties of learning the OSV language may have contributed to the rarity of this sentence structure in real-world languages (Derbyshire & Pullum, 1981). Accordingly, the high performance in SVO and SOV languages may be indicative of why these two word orders are the most common in the world (Greenberg, 1963).

Our study additionally intended to assess how familiarity with the word order of known languages affects the learning of new word orders, but due to the small sample size, no

conclusive results could be drawn. Of the 12 participants in the SOV group, only one participant was completely unfamiliar with any SOV languages. Furthermore, there were no significant relationships between SOV proficiency and performance in either 2AFC or Generation tasks. It is possible that we found no effect of language history on learning in the SOV group simply because the SOV word order was easy to learn; after all, performance in the SOV group – where participants had varying levels of familiarity with the word order – was as high as performance in the SVO group – where all participants had a high level of familiarity with the word order.

If a follow-up to this study were to be conducted, the participants could be selected in such a way that the number of monolinguals is nearly equivalent to that of multilinguals. Ideally, this study would be performed entirely with monolinguals of various native languages, so that no participant is familiar with more than one of the word orders being tested. Future language learning studies on the learning of a second language among adult populations may benefit from removing the conventional exposure phase from their methods, since no data can be collected about learning trajectory during this phase. The addition of a generation phase may also be beneficial, giving participants the freedom to generate the language themselves in order to assess how language production compares to language comprehension. It would also be constructive to conduct such trial-and-error methods with children of various developmental stages, since children learn language in a less explicit manner than adults, even in second-language learning (Ausubel, 1964).

Conclusion

In this study, we investigated the concurrent learning of semantics and grammar by taking the traditional AGL paradigm and including meaningful stimuli. Participants were capable of both comprehension and production of the artificial language, and these abilities were closely

tied together. Semantic knowledge was acquired simultaneously with grammatical knowledge, although participants' ability to comprehend the two aspects of language differed from their ability to produce sentences using that knowledge. Our results suggest that a language learning paradigm using trial and error as the primary means of learning can lead to rapid and effective learning, and perhaps this paradigm could be an effective way to teach natural languages as well.

References

- Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1), 67–82. <https://doi.org/10.1016/j.jml.2009.09.005>
- Ausubel, D. P. (1964). Adults versus children in second-language learning: Psychological considerations. *The Modern Language Journal*, 48(7), 420–424. <https://doi.org/10.1111/j.1540-4781.1964.tb04523.x>
- Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., & Christiansen, M. H. (2015). Factors influencing sensitivity to lexical tone in an artificial language. *Studies in Second Language Acquisition*, 37(2), 335–357. <https://doi.org/10.1017/s0272263114000849>
- Comrie, B. (1981). *Language universals and linguistic typology syntax and morphology*. The University of Chicago Press.
- Derbyshire, D. C., & Pullum, G. K. (1981). Object-initial languages. *International Journal of American Linguistics*, 47(3), 192–214. <https://doi.org/10.1086/465689>
- Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, 6(2). <https://doi.org/10.1515/cllt.2010.006>
- Fedor, A., Varga, M., & Szathmáry, E. (2012). Semantics boosts syntax in artificial grammar learning tasks with recursion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 776–782. <https://doi.org/10.1037/a0026986>
- Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, 303(5656), 377–380. <https://doi.org/10.1126/science.1089401>
- Frinsel, F., Trecca, F., & Christiansen, M. H. (2020). The picture guessing game: The role of feedback in active artificial language learning. In *the 42nd Annual Conference of the*

- Cognitive Science Society (CogSci 2020)* (pp. 2813-2819). Cognitive Science Society.
- Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, & R. Golinkoff, (Eds.) *Action meets word: How children learn verbs* (pp. 544-564). Oxford University Press.
- Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, *440*(7088), 1204–1207.
<https://doi.org/10.1038/nature04675>
- Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088.
<https://doi.org/10.1177/0956797612463705>
- Goldberg, A., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 468-477.
- Greenberg, J. H. (1968). Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of Language* (pp. 73–113). MIT Press.
- Herbranson, W. T., & Shimp, C. P. (2003). “Artificial grammar learning” in pigeons: A preliminary analysis. *Animal Learning & Behavior*, *31*(1), 98–106.
<https://doi.org/10.3758/bf03195973>
- Hoen, M., Golembiowski, M., Guyot, E., Deprez, V., Caplan, D., & Dominey, P. F. (2003). Training with cognitive sequences improves syntactic comprehension in agrammatic aphasics. *NeuroReport*, 495–499. <https://doi.org/10.1097/00001756-200303030-00040>
- Huddleston, R., Pullum, G. K., & Reynolds, B. (2022). *A Student's Introduction to English Grammar*. Cambridge University Press.
- Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80. <https://doi.org/10.1126/science.283.5398.77>

- Öttl, B., Jäger, G., & Kaup, B. (2017). The role of simple semantics in the process of artificial grammar learning. *Journal of Psycholinguistic Research*, 46(5), 1285–1308.
<https://doi.org/10.1007/s10936-017-9494-y>
- Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298(5593), 604–607. <https://doi.org/10.1126/science.1072901>
- Poletiek, F. H., Monaghan, P., Van de Velde, M., & Bocanegra, B. R. (2021). The semantics-syntax interface: Learning grammatical categories and hierarchical syntactic structure through semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(7), 1141–1155. <https://doi.org/10.1037/xlm0001044>
- Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, 133(2), 227–244. <https://doi.org/10.1037/0033-2909.133.2.227>
- Rüsseler, J., Gerth, I., & Münte, T. F. (2006). Implicit learning is intact in adult developmental dyslexic readers: Evidence from the serial reaction time task and artificial grammar learning. *Journal of Clinical and Experimental Neuropsychology*, 28(5), 808–827.
<https://doi.org/10.1080/13803390591001007>
- Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel Statistical Learning by 12-month-old infants. *Infancy*, 4(2), 273–284.
https://doi.org/10.1207/s15327078in0402_07
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top Tamarin Monkeys. *Cognition*, 107(2), 479–500. <https://doi.org/10.1016/j.cognition.2007.10.010>

- Smith, J., Siegert, R. J., McDowall, J., & Abernethy, D. (2001). Preserved implicit learning on both the serial reaction time task and artificial grammar in patients with Parkinson's disease. *Brain and Cognition*, *45*(3), 378–391. <https://doi.org/10.1006/brcg.2001.1286>
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. <https://doi.org/10.1016/j.cognition.2007.06.010>
- Trecca, F., McCauley, S. M., Andersen, S. R., Bleses, D., Basbøll, H., Højen, A., Madsen, T. O., Ribu, I. S., & Christiansen, M. H. (2018). Segmentation of highly vocalic speech via statistical learning: Initial results from Danish, Norwegian, and English. *Language Learning*, *69*(1), 143–176. <https://doi.org/10.1111/lang.12325>
- Wilson, B., Haslam, L., Poletiek, F., & Petkov, C. I. (2020). Artificial language learning: Combining syntax and semantics. In *the 42nd Annual Conference of the Cognitive Science Society (CogSci 2020)* (pp. 2838-2839). Cognitive Science Society.
- Wilson, B., Slater, H., Kikuchi, Y., Milne, A. E., Marslen-Wilson, W. D., Smith, K., & Petkov, C. I. (2013). Auditory artificial grammar learning in macaque and marmoset monkeys. *Journal of Neuroscience*, *33*(48), 18825–18835. <https://doi.org/10.1523/jneurosci.2414-13.2013>

Tables and Figures

Figure 1.....	7
Table 1.....	8
Figure 2.....	14
Figure 3.....	17
Figure 4.....	18
Figure 5.....	22
Figure 6.....	23
Figure 7.....	26
Figure 8.....	27
Figure 9.....	27
Figure 10.....	29
Figure 11.....	30
Figure 12.....	31