CHAPTER 6

Rule learning in adults in the auditory domain using Lindenmayer grammars

The first half of this chapter was published as Geambaşu, A., Ravignani, A., & Levelt, C.C. (2016). Preliminary Experiments on Human Sensitivity to Rhythmic Structure in a Grammar with Recursive Self-Similarity. *Frontiers in Neuroscience, 10*(281).

The second half is in preparation as Geambaşu, A., Ravignani, A. Toron, L., & Levelt, C.C. (in prep). Rhythmic recursion? Human sensitivity to a Lindenmayer grammar with self-similar structure in a musical task.

The goal of this chapter was to build upon the work from the previous chapters which focus on simple XYX-type or Marcus rule learning, and to explore the possibility of using a grammar that was more complex yet more ecologically valid in its structure.

6.1 Abstract

We present the first rhythm detection experiment using a Lindenmayer grammar (or L-system grammar), a self-similar recursive grammar shown previously to be learnable by adults when using speech stimuli. In Experiments 1 and 2, we presented adult learners with a passive exposure to the grammar, composed of two different drum sounds. Participants were then asked to identify whether test items followed the same grammar or not. Results of these first experiments

showed that learners were unable to correctly accept or reject grammatical and ungrammatical strings at the group level, although five (of 40) participants were able to do so with detailed instructions before the exposure phase. These results could have been due to the fact that ungrammatical test items could have also been generated by an L-system grammar themselves, making them difficult to distinguish from the grammatical strings. In Experiment 3 we used ungrammatical strings that could not have been generated by an L-system grammar. In addition we tested participants in two different paradigms: both a 2-alternative forced choice (2AFC) and a Yes/No task. After a brief exposure phase, we found that participants at the group level were sensitive to the exposure grammar and capable of distinguishing the grammatical and ungrammatical test strings above chance level in both tasks. However, the results were not robust and did not hold up in a more stringent statistical model. Hence, on the one hand we found modest evidence of participants' sensitivity to a very complex L-system grammar in a non-linguistic, potentially musical domain. On the other hand, our results were not robust. We discuss the discrepancy within our results and with the previous literature using L-systems in the linguistic domain. Furthermore, we propose directions for future language and music cognition research using L-system grammars.

6.2 Introduction

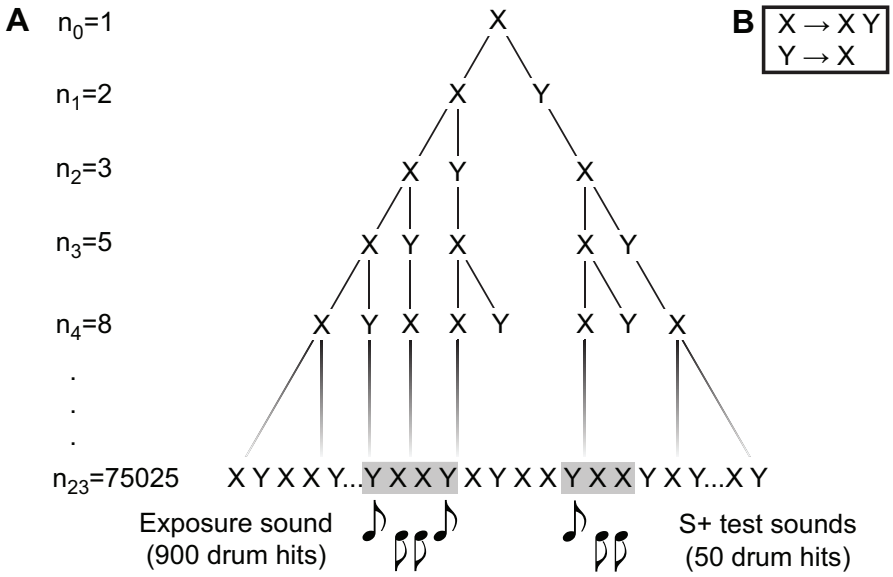
Processing of hierarchical structures has been proposed as a uniquely human ability, a hallmark of the linguistic system that distinguishes human language from animal communication systems (Hauser, Chomsky, & Fitch, 2002; Martins, 2012). Recursion is often considered the pinnacle of human-specific hierarchical structures (Hauser et al., 2002). Artificial Grammar Learning experiments have shown that adult participants are able to learn the context-free grammar A^nB^n, whose generation requires hierarchical rules, even without the need for semantic information (Lai & Poletiek, 2013). Parsing and generalizing grammars like A^nB^n requires detection that a structure, e.g., AB, is embedded between elements of another structure, e.g., A...B. Other species have not been shown unequivocally to be able to learn on the basis of the center-embedding principle required of A^nB^n (rather than using other strategies; Corballis, 2007; van Heijningen, de Visser, Zuidema, & ten Cate, 2009; Beckers, Bolhuis, Okanoya, & Berwick, 2012; Poletiek, Fitz, & Bocanegra, 2016; Ravignani, Westphal-Fitch, Aust, Schlumpp, & Fitch, 2015), which is taken as evidence that processing of recursion is a human-specific capacity.

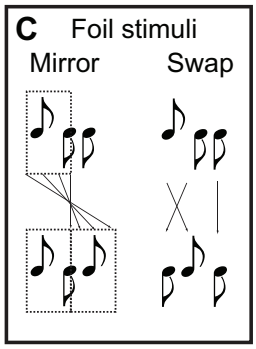
Yet to what extent learning of an A^nB^n grammar can be taken as evidence for processing recursive information at all is debated. Some researchers argue that human participants could in fact use simpler strategies, such as counting and matching the number of As and Bs in a test sequence (Hochmann et al., 2008; Zimmerer, Cowell, & Varley, 2011), while others argue that despite different strategies, the same core operations are nonetheless necessary

(Fitch & Friederici, 2012; Fitch, 2014). Saddy (2009) proposed that a more suitable grammar for the investigation of recursive processing may be Lindenmayer grammars, or L-systems. Uriagereka, Reggia, and Wilkinson (2013) have proposed that these grammars are suitable for between-species comparative work because they generate utterances that can be infinitely long and produce a "rhythm" when recognized. L-systems were first proposed by Lindenmayer to describe algae cell growth (Lindenmayer, 1968; Lindenmayer & Rozenberg, 1972) and have since been used to describe and recognize different plant structures (Samal et al., 1994). L-systems have rewrite rules that occur in parallel and have no terminal symbol, indicating that they can produce infinite sequences (Figure 6.1A). Because of their hierarchical structure and recursive properties, they are an interesting grammar to use in testing recursive processing. In her dissertation, Shirley (2014) began to explore the learnability of Fibonacci grammars, a subgroup of L-systems, that at each iteration produce sequences with lengths corresponding to Fibonacci numbers. She found that after a 3-min training with a Fibonacci grammar composed of syllables *bi* and *ba*, participants were able to correctly accept grammatical 10-s-long structures, and correctly reject ungrammatical ones. However, how participants processed the stimuli in Shirley's task is not clear yet. A possible rhythm-based strategy may have been used by participants to recognize a pattern in sounds generated by recursive branching, using rhythmic structure, i.e., how durational events are grouped and perceived hierarchically based on their relative accentuation. When presented with sequences of acoustic events occurring at constant time intervals (i.e., isochronous, as in Shirley, 2014), humans tend to group these events. Grouping often occurs when events are differentially accented, that is, marked by differing pitch or intensity (e.g., strong-weak-weak; Hay & Diehl, 2007).

The detection of a specific rhythmic pattern might be the mechanism participants draw upon to detect recursive structures such as those tested here. Syllables in Shirley (2014) differed by their vowel quality, with possibly some non-systematic variation in fundamental frequency and intensity. If detection strategies based on rhythmic features were used to learn Shirley's grammars, participant tested with percussion sounds (enhancing the recursive rhythmical structure of the stimuli) instead of speech syllables should show similarly high or even better performance, as the non-temporal rhythmic cues (intensity or pitch accentuation) would be enhanced, while violations in interstimulus intervals would disrupt the rhythmic detection strategy and hence grammar recognition; Shirley, 2014).

Can a complex pattern, recursively and hierarchically organized according to an L-system, be learned on the basis of a rhythmical strategy? We tested this hypothesis by enhancing the rhythmic quality of the sequences by using drum sounds differing in pitch and intensity, instead of syllables. This work thus constitutes the first study on rhythm perception using L-systems . We conducted two experiments (Figure 6.1D), each with two conditions (two types of





D

Experiment 1 limited instructions (20 participants)	**Mirror condition** S- were mirrored S+ (10 participants)	**Grammatical trial** Stimulus was S+ according to generative rule (18 trials)
		Ungrammatical mirror trial Stimulus was S- obtained by mirroring a grammatical string (18 trials)
	Swap condition S- were scrambled S+ (10 participants)	**Grammatical trial** Stimulus was S+ according to generative rule (18 trials)
		Ungrammatical swap trial Stimulus was S- obtained by scrambling a grammatical string (18 trials)
Experiment 2 specific instructions (20 participants)	**Mirror condition** S- were mirrored S+ (10 participants)	**Grammatical trial** Stimulus was S+ according to generative rule (18 trials)
		Ungrammatical mirror trial Stimulus was S- obtained by mirroring a grammatical string (18 trials)
	Swap condition S- were scrambled S+ (10 participants)	**Grammatical trial** Stimulus was S+ according to generative rule (18 trials)
		Ungrammatical swap trial Stimulus was S- obtained by scrambling a grammatical string (18 trials)

Figure 6.1: A derivation of the target Fibonacci grammar at the first four iterations and at the final 23rd iteration used to generate the exposure and test stimuli (A), the rewrite rules of the grammar (B), the makeup of the two foil grammars (C), and an overview of the two experiments reported with their two respective foil test conditions (D). We use upward and downward note stems to differentiate between the two drum sounds.

foil grammars) to evaluate the learnability of the L-system grammars. Between our two experiments, we also varied instructions, to further explore whether the method of presenting the exposure stimuli had an effect on learning ability. Based on previous work by Saddy (2009) and Shirley (2014) we expected that participants would pick up on the rhythmic nature of the structures, and be able to discriminate grammatical from ungrammatical strings. Our results indicate that for the majority of our participants, rhythm alone may not be enough to learn this type of grammar; musical background, age, instruction, and the specific types of foil grammars may all be contributing factors.

6.3 Experiments 1 and 2

6.3.1 Methods and materials

Two experiments were conducted, using Fibonacci grammars similar to those used in Saddy (2009) and Shirley (2014). The experiments consisted of an exposure phase and a test phase. During the exposure phase, participants passively listened to a sequence of kick and snare drum sounds following a Fibonacci grammar. During the subsequent test phase, participants were asked to indicate whether the test item (composed of the same kick and snare sounds) corresponded to the grammar from the listening phase, and to rate their certainty. The two experiments (Experiments 1 and 2) differed only in the detail of instruction given to participants. Instructions in Experiment 2 were more detailed than those in Experiment 1 (see Procedure). Each of the experiments consisted of two conditions (Mirror and Swap), in which each of the ungrammatical test items differed from the target Fibonacci grammar in different ways (see Stimuli).

Participants

Forty students (nine males; age range 18–32, M = 22, SD = 3.05) from Leiden University participated, N = 20 in Experiment 1 and N = 20 in Experiment 2. Participants were recruited via the SONA participant recruitment website of Leiden University. None of the participants had hearing problems or were dyslexic. Participants had various linguistic backgrounds, with all participants speaking at least one foreign language. They also had varying degrees of musical experience. The study was approved by the Ethical Committee of the Faculty of Social Sciences at Leiden University. Participants signed an informed consent form before taking part and were fully debriefed on the intention of the study upon completion of the experiment. They received course credits or monetary compensation for participating.

Stimuli

The Fibonacci sequences were made of simple drum sounds: a kick (average intensity 78 dB; sound X) and a snare (average intensity 66 dB, sound Y), each 200 ms in duration. See Figure Figure 6.1B for the Fibonacci grammar's rewrite rules.

An exposure string was created using a series of custom-written Python scripts, which created a large iteration of the Fibonacci grammar (n = 23, resulting in a 75025-element-long string). From this initial sequence, a 900-element (3-min-long) sequence was extracted and used for the habituation phase. Grammatical test items (50 elements, 10 s long) were extracted from the remaining sequence such that each grammatical string was unique.

Two modifications of the Fibonacci sequences were used as foil grammars. The first will be referred to as a Swap sequence. A Swap sequence consisted of a sequence taken from the remainder of the initial 75025-long sequence, in which a randomly-selected X and an adjacent Y from the string were switched, subject to the constraints that the swap would (i) produce a different string and (ii) not introduce an easily-detectable YY bigram (Figure 6.1C). For example, if the Fibonacci iteration n = 3 is XYXXY (see Figure 6.1A), its corresponding Swap sequence may be XXYXY. The second foil sequence will be referred to as a Mirror sequence. A Mirror sequence consisted of the Fibonacci sequence that was cut in half; this first half of the sequence was mirrored and replaced the original second half (Figure 6.1C). For example, if the Fibonacci iteration n = 5 is XYXXYXYXXYXXY (Figure 6.1A), its corresponding Mirror sequence would be: XYXXYX**Y**XYXXYX, where the seventh element (Y, bold) is treated as the point of mirroring. In order to avoid introducing more than two repetitions of the X element, or more than one repetition of the Y element, the point of mirroring varied by sequence, and thus mirror sequences could be either 50 or 51 elements long.

The composition of the foil grammars ensured that they never occurred in the habituation sequence, nor could they have ever occurred in any shorter iterations of the Fibonacci grammar. They also ensured that the grammatical and ungrammatical items were as similar as possible with respect to their local (element adjacency) and global (distribution of Xs and Ys) properties, thus preventing participants from solving the task by using simpler methods such as counting.

Materials

The experiments were conducted on a computer running Windows 7, with a 17-inch monitor (refresh rate: 60 Hz; resolution: 1280 x 1024 pixels). Participants sat ±50 cm from the screen in a quiet room and listened to the stimuli via headphones (Sennheiser HD 201). The experiment was programmed and run in Praat (Boersma and Weenink, 2014) and participant responses were registered via mouse clicks.

Procedure

In Experiment 1, participants were first presented with the following instruction: *You will now hear a 3-min-long rhythmic sequence. Listen carefully. When the sounds stop, press the spacebar to proceed to the test phase.* Participants in Experiment 2 were presented with more specific instructions: *You will now hear a 3-min long rhythmic pattern. Listen carefully. You will have to distinguish between this pattern and another pattern in the test phase. When the sounds stop, press the spacebar to proceed to the test phase.*

Within each experiment (Figure 6.1D), an equal number of participants was randomly assigned to the Mirror condition or the Swap condition (n = 10 per condition per experiment). In both conditions, the participants listened to the same L-system exposure sequence for 3 min. During the exposure phase the display was gray and showed a black fixation cross. After the exposure phase, the testing phase began. Participants were then presented with the following instructions: *The test phase will now begin. You will hear 36 test sounds. For every sound, listen carefully and indicate whether it follows the same rhythm as during the listening phase. Rate your certainty on a scale of 1 to 5. 1 = definitely no; 2 = probably no; 3 = not sure; 4 = probably yes; 5 = definitely yes. Only answer when the sound has finished playing.* During the test phase, participants in both the Mirror and the Swap condition were tested on their ability to discriminate between 10-s-long grammatical L-system sequences and ungrammatical sequences (Mirror or Swap sequences, depending on condition). In both conditions, they were instructed to indicate whether the sequences they heard followed the same rhythm as the sequences they had heard during the listening phase. The instructions appearing on the screen during playback of each test item were as follows: *Does this sound follow the same rhythm as in the listening phase? How sure are you?* Participants could then answer by clicking on one of two boxes with the words YES or NO. For their sureness response, they clicked on one of five boxes with numerals 1 (definitely no) through 5 (definitely yes).

Upon completion of the experiment, participants filled in a questionnaire, which inquired about their sex, age, hearing, dyslexia, languages spoken, handedness, musical training, and education level and background. They were subsequently debriefed on the purpose of the study and any questions they had were answered.

6.3.2 Descriptive statistics and results

There are two types of correct answers, namely a correct acceptance of a grammatical L-system sequence, and a correct rejection of an ungrammatical foil sequence. Thus, we analyzed correct responses both overall and comparing acceptances and rejections.

At the group level, when pooling across participants, the number of correct responses was at chance for each of the four groups (1 sample t-test, all t $<$

1.8, all $p > 0.12$). For each experiment and in each condition, performance did not differ between correct acceptances of grammatical and correct rejections of ungrammatical stimuli (paired samples t-test, all $|t| < 1.7$, all $p > 0.13$); also reaction times did not differ (all $t < 0.71$, all $p > 0.49$).

For each of the four groups (see Figure 6.2), we did a Spearman correlation (uncorrected) between % correct responses and:

- Median reaction time (correlations between –0.03 and 0.12, all $p > 0.72$)
- Age of participant (correlations between –0.18 and 0.38, all $p > 0.27$)
- Certainty of response (correlations between –0.16 and 0.43, all $p > 0.21$)
- Musical training (correlations between –0.11 and 0.39, all $p > 0.26$)
- Sex (correlations between –0.41 and 0.56, all $p > 0.08$).

Analyses of individual performances showed that five participants correctly classified stimuli above chance. A Fisher exact test revealed that each of these five participants significantly more often than chance associated correct Fibonacci-grammatical stimuli as similar to the sequences heard in the exposure phase and foils as dissimilar to the sequences heard during the exposure phase (one-sided, all $p < 0.05$, all prior odds ratio using Maximum Likelihood Estimate > 4.0). These were participants numbers 30 and 31 (Experiment 2, Swap group) and numbers 22, 33, and 37 (Experiment 2, Mirror group). Interestingly, all these participants received detailed instructions. Moreover four out of five reported having musical training (12 out of our 40 participants reported musical training).

6.3.3 Discussion

Our experiments did not show that, at the group level, participants were able to learn the Fibonacci grammars and discriminate them from either the Mirror or Swap foil grammars. At the individual level however, there were five participants in Experiment 2 who correctly identified grammatical and ungrammatical strings above chance level, suggesting that with specific instructions participants may be able to discriminate the grammatical from ungrammatical strings. Of those who did perform above chance level, most had received musical training, adding weight to the argument that rhythm perception may be involved in learning this type of grammar. However, it is probable that these participants performed well purely by chance. The question thus remains as to why most of our participants were not able to discriminate grammatical and ungrammatical strings, while the participants in Shirley (2014) were able to do this.

The very limited proficiency our participants achieved may be due to the fact that their memory trace for the exposure stimuli faded during the course of 36 test trials. Yet while there was no re-exposure phase per se during the presentation of the test sounds, the L-system test sequences being taken directly

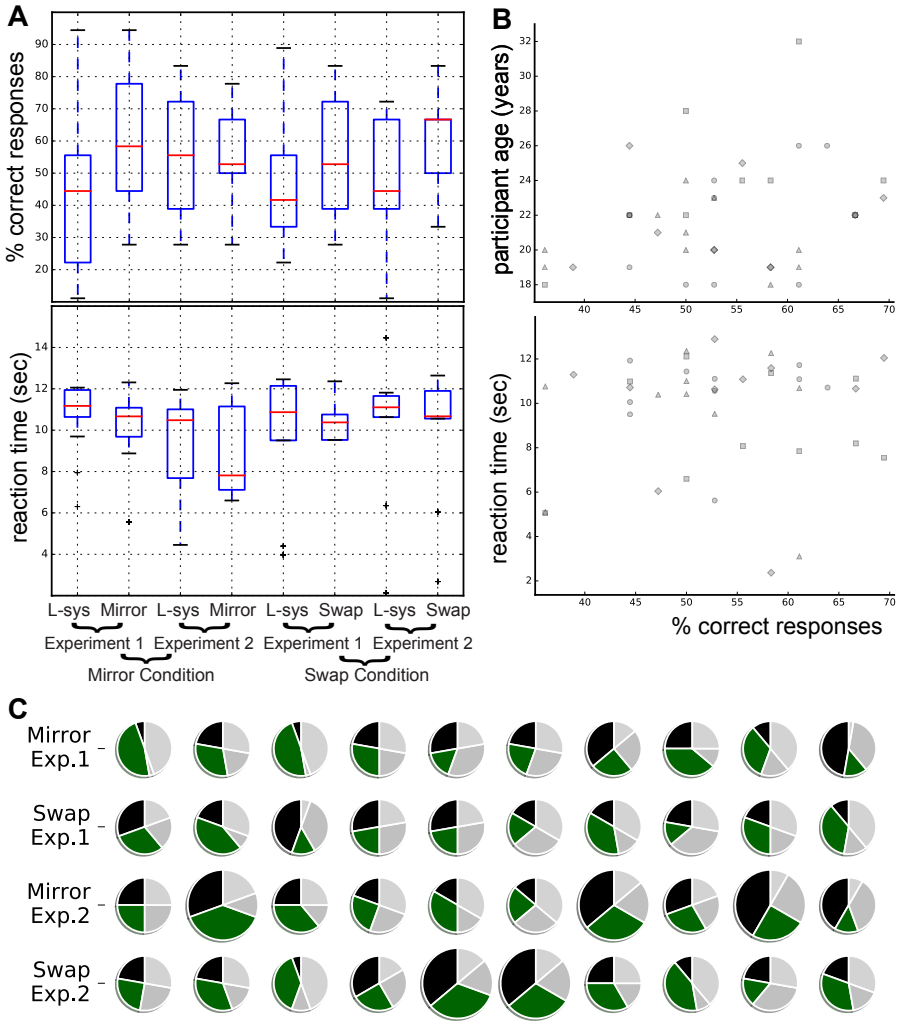


Figure 6.2: Summary of participants' performance at group (A) and individual (B,C) level. (A) Boxplot of percentage correct responses by experimental condition (Mirror vs. Swap), experiment number (limited vs. detailed instructions), and stimulus type (L-sys denotes a correct acceptance of a grammatical stimulus and Swap or Mirror denotes a correct rejection of an ungrammatical stimulus). (B) Individual % of correct responses is plotted against participant age and reaction time. Marker shapes denote experimental groups and conditions: mirror group without (circle) and with (square) specific instructions; swap group without (triangle) and with (diamond) specific instructions. (C) For each experiment, condition and participant, correct (black and green) and incorrect (silver and light gray) acceptances/rejection of grammatical and ungrammatical stimuli. Larger pies denote the five participants showing significance at an individual level.

from the L-system exposure sequence might be considered both a test and re-exposure. More important to their failure to perform was likely the fact that the foil grammars were too similar to the target grammar to be discriminated. While our exposure grammars were similar to those used in Saddy (2009) and Shirley (2014), our foil grammars differed in that ours did not include repetitions of both Xs and Ys, and thus could not be discriminated using repetition detection. By making the difference between target and foil grammar more subtle to avoid this method of discrimination, it might be that some of our foils were substrings of the Fibonacci-grammatical space, generated by one of the infinite iterations of the rewrite rules (Krivochen and Saddy, personal communication). This would have made discrimination between the target and foils more difficult in our experiment than in the experiments by Saddy (2009) and Shirley (2014), in which foils were part of the L-system space but not Fibonacci-grammatical. We can therefore not conclude whether or not participants are able to learn a Fibonacci grammar when presented with musical sounds. In future research, in order to be able to draw conclusions about whether musical rhythm differs from linguistic rhythm, and whether participants are able to use some sort of rhythmic structure to learn Fibonacci grammars (rather than surface properties of the stimuli) foil grammars should be calibrated to an optimal tradeoff between the structural properties of Shirley's foils and the surface properties of those used here. In addition, a different paradigm, such as Serial Reaction Time or EEG, may help illuminate what cues in the sequence participants attended to and at which point they detect an error.

In addition, several important points for consideration in future experiments are raised by our results. First, the individuals who performed above chance in correctly identifying grammatical and ungrammatical sequences, all took part in Experiment 2, where instructions were more specific than in Experiment 1. Instructions in Experiment 2 were also more in line with Shirley's instructions, letting participants know before training that they would later have to judge the correspondence between the test items and the exposure sounds. Our instructions did, however, differ from Shirley's in that Shirley used the word "language rule" whereas in our experiments, the term "rhythmic pattern" was used in order to potentially push participants even further in focusing on the rhythm of the sequences. The different terms may prime participants to listen to and learn about the same exposure grammars in different ways. Future experiments should thus take instruction into account as a factor[1]. Furthermore, another factor that should be taken into account and balanced in the future is age of participants; although not significant in the statistical analysis, older participants may perform better on this type of rhythm detection task (Figure 6.2B).

Taking into account the important difference in foil grammars between our experiments and those reported in Shirley (2014), we hypothesize that when given a complex grammar as foil that is not part of the Fibonacci grammatical

[1]See also chapter 5 of this thesis.

space, participants would be able to draw upon rhythmic detection abilities to accurately accept grammatical and reject ungrammatical sequences. Success of some individuals on our potentially more difficult task (as compared to Shirley's) already points in this direction.

6.4 Experiment 3

6.4.1 Methods and materials

In Experiment 3 we explore this possibility by using a new grammar as a foil which could not have been part of the Fibonacci grammatical space. We tested adult participants in two tasks commonly used in artificial grammar learning experiments: the two-alternative forced choice task (2AFC) and a yes/no judgment task (Yes/No). In this experiment, participants performed one of the two tasks first. In this experiment, we expected to see evidence of discrimination between grammatical and non-grammatical test items. We also expected to see a learning effect as manifested by an improvement from the first task to the second task, independent of what the first task was.

6.4.2 Participants

Participants were university students, of Dutch and international origin, recruited via the Leiden University Research Participation portal (SONA).

Participants were blind to the nature of the experiment before participating: they were told only that they were taking part in a task meant to test how people perceive rhythm. Upon completion of the experiment, participants filled in a questionnaire, which inquired about their sex, age, hearing, dyslexia, languages spoken, handedness, musical training, and education background and level. They were subsequently debriefed on the purpose of the study, and any remaining questions they had were answered.

The experiments were approved by the ethical committee of the Faculty of Social Sciences of Leiden University. Participants received course credit or monetary compensation for participation.

We tested 34 participants, two of whom were excluded from analysis, one due to technical error, and the other due to dyslexia and hearing deficits. The results are based on the experimental data of the remaining 32 participants (16 per order; in each order 11 females and 5 males; in each order age range 18-26, M_{2AFC}=22.31, SD_{2AFC}=2.41, M_{YN}=22.25, SD_{YN}=2.50). There were 10 musicians in the 2AFC-first order and 8 musicians in the Yes/No-first order.

6.4.3 Stimuli

The Fibonacci sequences were identical to those used in Geambașu, Ravignani, and Levelt (2016). They were made of simple drum sounds: a kick (average

intensity 78 dB, average pitch 108 Hz; sound X) and a snare (average intensity 66 dB, average pitch 168 Hz; sound Y), each 200 ms. in duration. The items followed a Fibonacci rewrite rule resulting in an 18-item long sequence (see Figure 6.1 for rewrite rules), which was then repeated 23 times to form the familiarization grammar. This resulted in a sequence of 75025 items, three minutes in duration. The Fibonacci grammatical test items were extracted from this familiarization stream in such a way that each test string was unique. The test items consisted of 50 elements and were 10 seconds in duration.

The foil test grammar was a regular grammar composed of the 18-item long sequence XYXYXXYXXYXYXXYXXY repeated three times. This created a string of 54 sounds, 10 seconds in duration. Both grammatical and foil test strings could begin or end with either an X or a Y element. The composition of the foil grammars ensured that they never occurred in the habituation sequence, nor could they have ever occurred in any shorter iterations of the Fibonacci grammar. They also ensured that the grammatical and ungrammatical items were as similar as possible with respect to their local (element adjacency) and global (distribution of Xs and Ys) properties, thus preventing participants from solving the task by using simpler methods, such as counting. However, as opposed to the foils used in Geambaşu et al. (2016), the foil grammar used in the present work can be characterized as a regular grammar and could therefore not be part of the Fibonacci grammatical space. Thus, if participants were memorizing substrings of the L-system and comparing them with the foil, they would fail discriminating the two types of sequences. However if participants internalized a rule to generate the grammatical strings, they would not accept the foil grammar.

6.4.4 Materials

As in the previous experiments, the experiment was programmed and run in Praat version 5.4 (Boersma & Weenink, 2014) and was conducted on a computer running Windows 7, with a 17-inch monitor (refresh rate: 60Hz; resolution: 1280 x 1024 pixels). Participants sat approximately 50 cm from the screen in a quiet room and listened to the stimuli via headphones (Sennheiser HD 201). Participants responded by clicking boxes indicating their responses on the computer screen via mouse.

6.4.5 Procedure

In this procedure, participants were first familiarized with the target L-system grammar, then tested in one of the two testing paradigms, familiarized again, and finally tested with the other testing paradigm.

All participants saw the following instructions before either familiarization period: *You will hear a three-minute long rhythmic pattern. Listen carefully. You will have to distinguish between this pattern and another pattern in the test phase.*

An equal number of participants (n=16) was randomly assigned to be tested first with either the 2AFC test or with the Yes/No test. The conditions differed in their initial instructions to participants.

Before the 2AFC test phase, participants saw the following instructions: *The test phase will now begin. You will hear 18 pairs of test sequences. Each pair is separated by a 1 second silence. For every pair of sound sequences, listen carefully to both, and indicate which sequence follows the same rhythm as the listening phase: the first or the second one?*

Before the Yes/No test phase, participants saw the following instructions: *The test phase will now begin. You will hear 36 test sounds. For every sound, listen carefully and indicate whether it follows the same rhythm as during the listening phase by clicking "yes" or "no."*

In both testing conditions, the above instructions were followed by the following: *Rate your certainty on a scale of 1 to 5. 1 = very unsure / 2 = somewhat unsure / 3 = not sure / 4 = somewhat sure / 5 = very sure. Only answer when the sound has finished playing.*

6.4.6 Analysis of the data

Participants' responses in each case were recorded and our dependent variable per task was correctness of response. Participants' sureness scores were also analyzed. Frequency of correct responses were analyzed using a Chi-squared test. The proportion of correct scores as well as the proportion of participants' sureness in their responses were analyze using a binomial test. Finally, a logistic regression was performed to assess the predictive power of our independent variables of order and task on participants' performance.

6.4.7 Results

The data was analysed using RStudio 1.0.153 (RStudio Team, 2015).

Our data were divided along two dimensions, namely: order (which task was performed first), and task (which task was being performed at each response point).

We first tested the hypothesis that, at the group level, participants' performance would be better overall depending on which order they participated in. Order on its own did not influence participants' performance. While participants in the 2AFC-first order seemed to show more stable performance on both the 2AFC and the Yes/No tasks (Table 6.1), a Pearson's Chi-squared test with Yates' continuity correction also indicated that the difference in performance between the two orders was not significant [$\chi^2(1) = 0.087$, p=0.77].

We also tested whether there would be a difference in performance between the respective first and second task when the results were analyzed per order. We performed a Chi-squared test with Yates' continuity correction per order, testing whether the frequencies of correct answers within each order differed between tasks. There was no significant difference between tasks in correctness

in either the 2AFC-first order [$\chi^2(1, N=864) = 0.586$, $p = 0.44$] or in the Yes/No-first order [$\chi^2(1, N=864) = 1.538$, $p = 0.21$]. These results indicate that there was no effect of learning across the two tasks being performed, in either of the two orders.

However, we also tested whether order interacted with the task being performed. A Chi-squared test with Yates' continuity correction comparing frequencies of correct responses in the 2AFC task between the 2AFC-first order and the Yes/No-first order showed a significant difference between orders [$\chi^2(1, N=576) = 4.475$, $p = 0.034$], with participants performing the 2AFC task better, on average, if it preceded the Yes/No task, than if it followed the Yes/No task. When performing the Yes/No task first, however, the difference between correct and incorrect responses was not significantly different between the two orders [$\chi^2(1, N=1152) = 0.177$, $p = 0.674$] showing no evidence of an effect of order on this task.

Table 6.1: Percentage of correct responses, overall and in selected subgroups

Overall	59.26%
2AFC first condition	61.23%
• *2AFC task*	• 63.19%
• *Y/N task*	• 60.24%
Yes/No first condition	57.29%
• *Y/N task*	• 58.85%
• *2AFC task*	• 54.17%
2AFC task	58.68%
Yes/No task	59.55%

Binomial test

When analyzing the results of the tasks individually, we find that performance in each task is above chance (which is assumed as the 0.5 probability of choosing the correct response in either task). In order to exclude that this performance is actually part of a population that performs at the 0.5 level, we performed one-tailed (greater) binomial test. Across tasks and independent of order, the binomial test indicated that the frequency of correct responses of 59.26% was significantly higher than the chance level of 0.5, $p < 0.01$ (one-sided). For the Yes/No task, the frequency of correct responses of 59.55% was also significantly higher than 0.5, as it was for the 2AFC task (with a frequency of 58.68%; both $p < 0.01$; see Figure 6.3). These results are independent of whether each task was performed first or second.

Because we have a large total sample size of 1728 trials, we also calculated

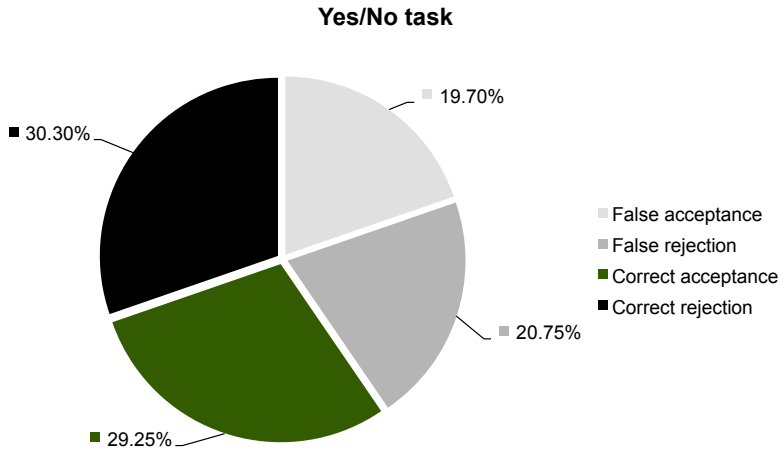


Figure 6.3: Correctness by Category in the Yes/No task. Correct acceptance and correct rejection were both respectively more common than false acceptance or false rejection.

the relative risk of each of these measures. We calculated the relative risk as the proportion of the probability of success against 0.5. The values per order and per task are shown in Table 6.2. See also Figure 6.4.

Table 6.2: Outcomes of binomial tests, p = 0.5, one-tailed (upper)

	Probability of success	p-value	Relative risk
2AFC first condition	0.612	< .001	1.225
Yes/No first condition	0.573	< .001	1.146
2AFC task	0.587	< .001	1.174
Yes/No task	0.595	< .001	1.191
Overall	0.593	< .001	1.185

To give an indication of how certain participants were about their decision per trial, we analyzed the percentage of the sureness that participants had indicated per response. Overall, participants felt "somewhat sure" or "very sure" about 41.4% of their responses, "not sure" (neutral) about 31.2% of their responses, and "somewhat unsure" or "very unsure" about 27.4% of their responses, indicating more of a tendency towards certainty. Their certainty supports the finding that they respond more correctly than expected by chance.

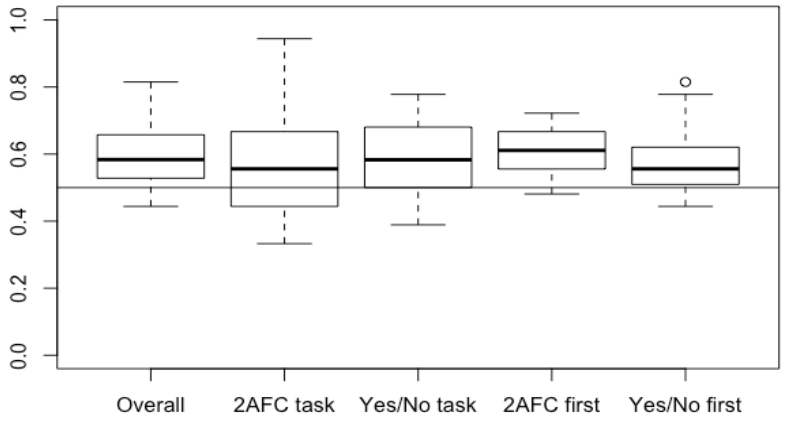


Figure 6.4: Proportion correct of participants of all data, split across tasks, and split across conditions (line shows 0.5 chance level).

Logistic regression

In order to have an overview of how our independent variables were predicting the correctness of participants' responses, we performed a logistic regression of correctness against task (Yes/No or 2AFC), order (Yes/No First or 2AFC First) and stimulus type (grammatical or ungrammatical) of all data, as well as of correctness against order and stimulus type of both individual tasks (with a binomial distribution). We performed the logistic regression in RStudio, using the generalized linear model function, with family specified as binomial and link as logit. Comparisons between models were done using a chi-square test comparing the full model against the intercept-only model. When looking at the overall model of correctness predicted by task, order, and stimulus type, none of these factors seem to significantly contribute, nor did we find any interaction effects, N=1728, all $p>.13$.

6.4.8 Discussion

The results of this experiment indicate that participants were able to learn the exposure grammar and correctly categorize the test items according to what they had learned. We thus find that when the foil test items cannot be a part of the L-systems space but does share surface properties with the L-system grammar, participants are able to distinguish the grammatical and ungrammatical stimuli.

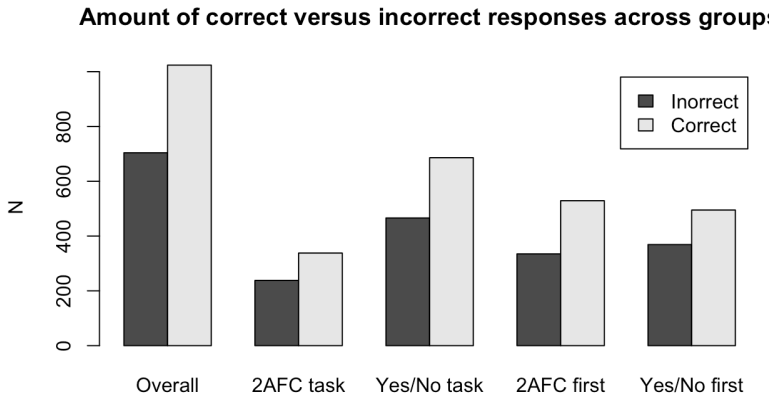


Figure 6.5: Number of correct versus incorrect responses per task and per condition. In all cases, participants provided more correct responses than incorrect responses.

6.5 General Discussion

Across three experiments, we tested participants' capacity to discriminate between drumming sequences built according to a recursive L-system (grammatical stimuli) and foil sequences (ungrammatical stimuli). In Experiments 1 and 2, when the foil grammars presented to the participants were potentially part of the Fibonacci grammatical space, participants were unable to discriminate them from the exposure grammar. While our first two experiments can be considered an exploratory pilot, showing the difficulties of finding an appropriate foil grammar four our AGL study, our Experiment 3 indicate that when an appropriate foil grammar is used (i.e., one not part of the L-systems space but sharing surface properties with the grammatical stimuli), participants show signs of being able to learn from the complex exposure grammar. In Experiment 3, we found that participants were partly sensitive to the exposure grammar, overall capable to distinguish the grammatical and ungrammatical stimuli. Binomial tests showed overall correct categorization of test items according to grammatical items presented during exposure at a higher rate than expected by chance, in both a Yes/No task and a 2AFC task. However, a more structured logit model did not find any significant effect. These partly contradictory statistical outcomes suggest that, while participants may have picked up some of the grammatical regularities they were exposed to, they did this neither strongly nor robustly.

There are several potential reasons for these partly contradictory results. First, our study may be underpowered. Considering the complexity of the grammar, a larger sample size could have helped detecting a small effect. Second, learning may have been occurring over trials within one or both tasks, leading to significance depending on the statistical test employed. Finally, there may

have been a high inter-individual variability: individual differences are common sources of variance in grammar learning experiments.

Nevertheless, there was some evidence of sensitivity of participants to the target L-system grammar. The grammar is too complex for participants to be able to explicitly state what rule generates it. However, as we controlled for surface similarities between the grammatical and foil stimuli, participants' performance indicates that they formed an implicit sensitivity to regularities found on a level deeper than simply the surface level. In our previous work, when the foil grammars presented to the participants were potentially part of the Fibonacci grammatical space, participants were unable to discriminate them from the exposure grammar. However, in the present work, where the foil grammar could not have been a part of the Fibonacci grammatical space but still shared surface properties with the exposure grammar, participants were able to pick out this grammar as different and ungrammatical, in both a Yes/No task and a 2AFC task.

Yet there are some caveats that may apply to these results. First, we must note that the foil grammar may be predictable in that if the sequence XXY is present it will always be followed by the same sequence (i.e., XXYXXY is highly likely). The L-grammar lacks this regularity. It may be that participants were sensitive not to the regularity of the L-system grammar, but to the regularity in the foil system which was not present in the L-system. This type of processing has been found in AGL tasks with birds: in Chen and ten Cate (2015) zebra finches seem to learn a rule based on the presence or absence of bigrams rather than the whole structure. In human AGL, we find evidence of simpler mechanisms as well, such as a deviant detection mechanism that allows participants to successfully complete the task without learning about the whole of the structure. This may be a learning mechanism that helps learners in early stages of learning from a complex input stream (van der Kant, 2015).

While we did not find that task on its own had an effect on learning, there was evidence of better outcomes when participants were performing the 2AFC task when it was the first task being performed. This may be because although both tasks are considered recognition tasks, the Yes/No task may be characterized as an identification task in which participants must categorize a single stimulus as correct or incorrect, while the 2AFC task is both an identification and discrimination task in which the two stimuli must first be differentiated and a correct one must be chosen. As such, they tap into different recognition mechanisms. The 2AFC task performed immediately after exposure may be better suited to tap into sensitivities that participants may have gained during the exposure phase (Jang, Wixted, & Huber, 2009).

Finally, the results seen here are undoubtedly weaker than those found by Shirley (2014), where participants had a mean identification accuracy of L-system grammatical items in the range of 70 to 80% in a 2AFC task (Shirley, 2014, chapter 3). The discrepancy points to the facilitating role of speech stimuli for structure learning. Nevertheless, the fact that our participants show more

sensitivity to the L-systems grammars than would be predicted by chance when they are instantiated with non-speech, drum sound stimuli, indicates that L-systems have the potential to be a useful tool for AGL experiments, replacing simpler grammars across domains.

Improvements to these experiments can be made to gain a better understanding of how participants process these grammars in real time and at which points they may struggle. To this end, tasks that incorporate immediate reaction times, such as Serial Reaction Time tasks, can be employed. We have already started working in this direction with a simultaneous tapping experiment (Ravignani, Geambasu, Minnema, & Levelt, in prep.). Such online measures should give us a better understanding of how participants process complex grammars, which we know they can eventually do.

6.6 Acknowledgments

We dedicate this work to the memory of Remko Scha. We thank Johanne Rauwenhoff for help collecting data in Experiments 1 and 2. We also thank Doug Saddy, Liz Shirley, and Diego Gabriel Krivochen for valuable discussion on the experiments. Andreea Geambaşu and Clara C. Levelt were supported by NWO Vrije Competitie grant 360.70.452 (to Clara C. Levelt). Andrea Ravignani was supported by ERC grants 283435 ABACUS (to Bart de Boer), 230604 SOMACCA (to W. Tecumseh Fitch) and ESF grant 5544 INFTY (to Andrea Ravignani).

6.7 Appendices

6.7.1 Appendix A

Overview of the data files and their formats

The raw data files from Experiment 1 are available at the figshare repository: https://figshare.com/s/83987b4a52906c87e115. The raw data is contained in the file "alldata.csv," which can be read by any text editor or Microsoft Excel. This file was obtained by merging all output files from individual participants (collected between Dec 5th, 2014 and Feb 26th, 2016), and adding additional information from questionnaire (e.g., musical training). Python scripts used for the analyses are available from the authors on request.

Variable names and coding (values in brackets):

- Experiment number: Experiment with limited (1) or detailed (2) instructions.
- Condition id: Participant was tested with Mirror (0) or Swap (1) stimuli.

- Condition name: alphanumeric string XY indicating the testing condition X and the experiment number Y.
- Participant: anonymized identifier for each participant (1,2,...,40).
- Trial number: number of trial in order of presentation (1,2,...,36).
- Stimulus type: test item was a string generated using an L-grammar (0) or a Swap/Mirror string (1).
- Response: Participant judged test stimulus to have the same (1) or a different (0) rhythm as those in the exposure.
- Correctness: participant chose the correct (1) or incorrect (0) response.
- Correctness by category: correct acceptance (1) or wrong rejection (2) of a string generated using an L-grammar; correct rejection (3), or wrong acceptance (4) of a string generated by swapping or mirroring elements (depending on the experimental group).
- Goodness: Whether a participant was very sure the sequence was correct (5) or very sure the sequence was incorrect (1) (1,...,5).
- RT: reaction time in seconds.
- Age: Years of age.
- Sex: Female (0) or male (1).
- Musical Training: Participant had some (1) or no (0) musical training.