

# Rhythmic Recursion? Human Sensitivity to a Lindenmayer Grammar with Self-similar Structure in a Musical Task

Andreea Geambaşu<sup>1,2</sup> , Laura Toron<sup>3</sup> , Andrea Ravignani<sup>4,5</sup>  
and Clara C. Levelt<sup>1,2</sup>

Music & Science  
Volume 3: 1–11

© The Author(s) 2020  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2059204320946615  
journals.sagepub.com/home/mns



## Abstract

Processing of recursion has been proposed as the foundation of human linguistic ability. Yet this ability may be shared with other domains, such as the musical or rhythmic domain. Lindenmayer grammars (L-systems) have been proposed as a recursive grammar for use in artificial grammar experiments to test recursive processing abilities, and previous work had shown that participants are able to learn such a grammar using linguistic stimuli (syllables). In the present work, we used two experimental paradigms (a yes/no task and a two-alternative forced choice) to test whether adult participants are able to learn a recursive Lindenmayer grammar composed of drum sounds. 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. While we found evidence of participants' sensitivity to a very complex L-system grammar in a non-linguistic, potentially musical domain, the 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 music cognition research using L-system grammars.

## Keywords

Artificial grammar learning, biomusicology, music and speech, music cognition, recursion, rhythm perception

Submission date: 7 August 2019; Acceptance date: 12 July 2020

## Introduction

Structure seems a core property of both language and music. Human adults have been shown to learn a context-free grammar  $A^nB^n$ , generated via hierarchical rules, in artificial grammar learning tasks, even when all semantic, linguistic, or musical information is absent (Lai & Poletiek, 2013). Recursion is a particular type of hierarchical structure, consisting of embedding one structure into a copy of itself, potentially infinitely many times (Martins, 2012). Some argue that the cognitive capacity to process recursive structures is uniquely human (Hauser et al., 2002). Several experiments have explicitly targeted recursion (e.g., Ferrigno et al., 2020; Martins, 2012; Martins & Fitch, 2014; Martins et al., 2016, 2017, 2020; Uddén et al., 2019), but it is still debated whether learning (hierarchical-like)  $A^nB^n$  grammars constitutes evidence for processing recursive information. While  $A^nB^n$  grammar requires that AB pairs are embedded recursively within other AB

pairs, resulting in strings such as  $A[AB]B$ ,  $A[A[AB]B]B$ , etc. (see Bahlmann et al., 2006, Figure 1 for a visualization of this), participants in artificial grammar learning tasks probing  $A^nB^n$  grammars might be able to solve such tasks via simpler mechanisms. One such shortcut could be counting whether or not strings contain an equal number of As

<sup>1</sup> Leiden University Centre for Linguistics, Leiden University, Leiden, The Netherlands

<sup>2</sup> Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands

<sup>3</sup> Radboud University, Nijmegen, The Netherlands

<sup>4</sup> Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium

<sup>5</sup> Comparative Bioacoustics Group, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

## Corresponding author:

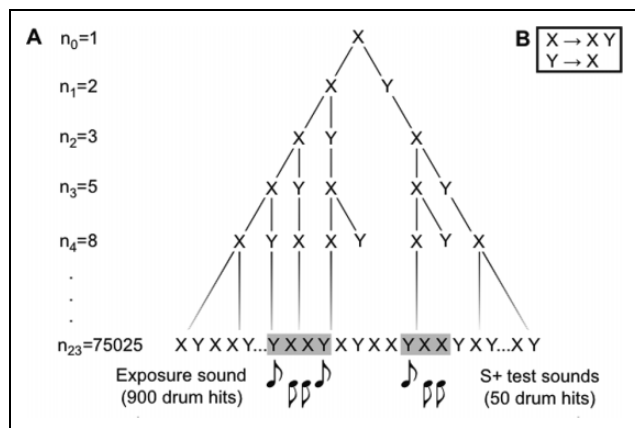
Andreea Geambaşu, Leiden University Centre for Linguistics, van Wijkplaats 4, 2311 BX Leiden, The Netherlands.

Email: a.geambasu@hum.leidenuniv.nl



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified

on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).



**Figure 1.** The Fibonacci grammar at the first four and final iteration used to generate the exposure and grammatical test sequences (A), and the rewrite rules of the grammar (B). Figure reproduced verbatim from Geambaşu et al., (2016), an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

and Bs (Hochmann et al., 2008; Zimmerer et al., 2011). In contrast, however, even if participants use such strategies because of simplicity, they may still possess core mechanisms that allow for hierarchical rule processing (Fitch, 2014; Fitch & Friederici, 2012). To what extent are humans and other animals sensitive to recursive properties instantiated in various stimuli (linguistic, visual, action, musical)? This remains an open question (with recent exciting developments, see e.g., Ferrigno et al., 2020). Surely, when addressing this question, it is necessary to employ artificial grammar stimuli which preempt the use of simpler mechanisms and strategies.

With few exceptions, hierarchy and recursion are usually tested in the linguistic domain. To better understand the role of recursion as a mechanism used in specific domains, here we focus on testing perception of music recursion. The concept of recursion in music is not new. Admittedly adopting different definitions of recursion (Martins, 2012), recursive procedures are often used to *generate* computer music (Loy & Abbott, 1985; Mazzola et al., 2016; Manaris & Brown, 2014; Prusinkiewicz et al., 1989; Yadegari, 1991). Likewise, ideas from recursion are employed to *analyze* the potential self-similar nature of music compositions (Gollin, 2008; Katz & Pesetsky, 2009; Lerdahl & Jackendoff, 1996; Losada, 2007; Mazzola et al., 2016; Murphy, 2007; Peck, 2004; Wooller et al., 2005). This work, while obviously relevant, cannot say much about *processing* and *cognition* of recursive musical patterns. Research showing that humans have some capacities to perceive hierarchical and recursive processes in music is much more relevant here (Koelsch et al., 2013; Martins et al., 2017); to our knowledge, this work is unfortunately still scarce and focused on melodic and harmonic properties of music. With our experiments, we aim to better characterize the human sensitivity to musical recursion in the rhythmic domain.

In studying the ability to process recursion, Lindenmayer grammars (L-systems; Lindenmayer, 1968) may be more appropriate than the more commonly-tested  $A^nB^n$  structures. Lindenmayer grammars have no terminals, meaning they are composed of rewrite rules that can generate infinitely long utterances. Moreover, some researchers have proposed that these grammars produce a “rhythmic” sensation in human listeners (Saddy, 2009; Shirley, 2014; Uriagereka et al., 2013), allowing for interesting cross-domain comparisons between music and language processing. In the speech domain, when participants were exposed to strings composed of speech syllables *bi* and *ba*, they could discriminate Fibonacci-grammatical utterances (a subgroup of L-systems) from non-grammatical ones (Saddy, 2009; Shirley, 2014). Yet, how participants process and learn these grammars is not clear. Whether they use language-specific, domain-general, or specifically musical mechanisms is an open question. In these experiments, a rhythm-based strategy could have been employed that draws upon metric structure, which allows for auditory stimuli to be grouped hierarchically based on differences in pitch or intensity. This would entail that even though musical stimuli were not explicitly used, participants perceived them implicitly as rhythmic due to their physical properties. In the present work, we aimed to disentangle these possibilities, by removing the linguistic aspect of the original work and specifically testing whether processing of this recursive grammar can be done in a non-linguistic domain, indeed on the basis of a rhythmic strategy. To this end, we enhanced the rhythmic quality of the L-systems output by generating auditory sequences composed of two different drum sounds rather than two syllables.

In previous work using the same stimuli, we were unable to show discrimination between an L-system grammar and a foil (2016), likely due to the fact that our foil strings could actually have been one of the possible L-system generated grammars (Diego Krivochen, person. commun.; Krivochen & Saddy, 2018). In the present work we therefore replaced our foil strings with ones that we assume did not belong to an L-system but nonetheless shared important surface properties with the target auditory sequences (see stimulus section below). This way, if participants learned the recursive properties of the familiarization stimuli, they would be able to discriminate between grammatical and non-grammatical test strings. Conversely, if participants only attended to the surface level properties of the test strings, they would not be able to easily discriminate the grammatical and non-grammatical strings, due to their similar surface forms.

We tested adult participants in two tasks, which are commonly used in artificial grammar learning experiments: a two-alternative forced choice task (2AFC) and a yes/no judgment task (Yes/No). Participants always performed both types of tasks, but they were evenly and randomly divided into two subgroups, each performing one of the two tasks first. Testing participants in two tasks allowed

us to investigate whether learning would occur between the first and second task, whether one type of task is better suited to show learning effects than the other, and whether performing one type of task before the other would enhance performance in the second (e.g., benefit in performing Yes/No first but not 2AFC first).

First, we hypothesized participants would be able to discriminate the target (grammatical) stimuli from the foil stimuli, as this ability had been previously shown with syllable strings (Shirley, 2014). Additionally, we also hypothesized that participants would have more correct responses in the second task, independently of which task was being performed second, as a result of gaining more experience with the target grammar. Finally, in order to test whether musicianship had any influence on the ability to recognize recursion in musical stimuli, we also balanced the number of musicians and non-musicians we tested. We hypothesized that musicians would be better able to recognize and distinguish regularities in music. On the other hand, if the ability to recognize strings containing recursion and distinguish them from those not containing recursion is a general human trait, we would expect to see above-chance performance from both musicians and non-musicians alike.

## Methods

### Participants

Participants were university students, of Dutch and international origin, recruited via the SONA Research Participation portal of Leiden University. Participants were blind to the purpose 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, participants were asked to fill in a background questionnaire which asked for age, sex, hearing problems, diagnosed dyslexia, handedness, a list and self-rating of known languages, their education background and level, and whether they had musical training (see supplementary material at [https://osf.io/s2f3h/?view\\_only=3191f5635f4b4dc48093ac36950f733f](https://osf.io/s2f3h/?view_only=3191f5635f4b4dc48093ac36950f733f)). Afterwards, they were given more details on the purpose of the experiment. Participation in our study was voluntary, and participants received a small remuneration or course credit for taking part. The experiment was approved by the ethics committee of the Faculty of Social Sciences of Leiden University.

We tested 34 participants, two of whom were excluded from analysis; one was excluded due to a technical error, and the other due to dyslexia and hearing deficits. The results are based on the experimental data of the remaining 32 participants. Sixteen participants took part in each of the two task orders (11 females and 5 males per task order; age range 18–26 per task order,  $M_{2AFC-first}=22.31$ ,  $SD_{2AFC-first}=2.41$ ,  $M_{YN-first}=22.25$ ,  $SD_{YN-first}=2.50$ ). As there are not enough similar studies to provide a meaningful power

analysis, we used the same number of participants as in a previous experiment (Geambaşu et al., 2016).

We tested participants with and without musical training. Musicianship was determined by self-reports on the participant background questionnaires. Participants in both testing orders indicated they had between one and 10 years of experience with a variety of instruments. There were 10 musicians in the 2AFC-first order (three of whom reported they were self-taught) and eight musicians in the Yes/No-first order (all of whom reported that their musicianship was a result of formal instruction).

### Stimuli

A series of Python scripts generated an initial Fibonacci-grammatical string, from which one familiarization string and 18 grammatical test strings were extracted. The Fibonacci-grammatical sequences were identical to those used in our previous work (Geambaşu et al., 2016). They were composed of two drum sounds, each 200 ms in duration (henceforth “elements”): 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). The items followed a Fibonacci rewrite rule (see Figure 1B for rewrite rules). The 23rd iteration of the grammar produced an “initial string” of 75,025 elements. From this initial string, we extracted a string of 900 contiguous X/Y elements (three minutes in duration) to use as the familiarization stream. We extracted the 18 unique test strings from the remainder of the initial string, each of them consisting of 50 contiguous X/Y elements (10 seconds in duration).

There were 18 pseudo-Fibonacci (pseudo-Fib) foil test strings, each selected from the initial 50-elements Fibonacci-grammatical string. For each of the foil test strings, a script selected a different 15-element-long sequence (for example, XYXYXXYXXYXXYXXY) from the initial string and repeated it four times, creating a string of 60 sounds. Of that 60-element-long string, our script selected the first 50 elements, leading to a 50-element-long sequence, such that test and foil strings had an identical number of elements and an equal duration while maintaining similar surface properties. Pseudo-Fib foil strings never occurred in the familiarization string in their entirety (a Python script checked whether our foil strings appeared anywhere in our initial string from which we extracted the familiarization and the grammatical test strings), nor could they have ever occurred in any smaller iterations (i.e.,  $n < 23$ ) generated by the Fibonacci grammar. Furthermore, both grammatical and foil test strings could begin or end with either an X or a Y element, and foils with two (or more) repetitions of Y and three (or more) repetitions of X were excluded as they could have never occurred in the test stimuli, and would have hence made discrimination extremely obvious. By creating these controlled pseudo-Fib foil strings, we aimed to ensure that the grammatical and ungrammatical strings were maximally

similar in their surface properties, including the number and distribution of Xs and Ys and their transitional probabilities. Each of these measures was important to control in order to prevent participants from being able to rely on simpler methods to solve the task (e.g., Ravignani et al., 2015; van Heijningen et al., 2009). All sound files used in the experiment can be accessed at [https://osf.io/s2f3h/?view\\_only=3191f5635f4b4dc48093ac36950f733](https://osf.io/s2f3h/?view_only=3191f5635f4b4dc48093ac36950f733). Sound files found at the link correspond to the dictionary found in Appendix A.

As opposed to the foils used in previous work (Geambaşu et al., 2016), the pseudo-Fib foil strings used in the present work would be accepted by a finite automaton,<sup>1</sup> and should therefore not be a part of the Fibonacci-grammatical space. Although there is no closed-form mathematical theorem proving this yet, in principle if a string is accepted by a finite automaton, it should not be Fibonacci-grammatical. In other words, the field of mathematics proving whether a particular Fibonacci grammar belongs to a particular subset of the Chomsky hierarchy (and vice versa) is still underdeveloped; there is still no clear roadmap of all possible strings that are not Fibonacci. Our choice of control stimuli tried to strike a balance between the foil stimuli's surface similarity to the test stimuli and their likely belonging to a different grammar than test stimuli. Because of that, according to our choice of foil stimuli, if participants were memorizing substrings of the Fibonacci-grammatical familiarization string and comparing them with the foil strings, they would likely fail at discriminating the two types of sequences. However, if participants internalized a rule to generate the grammatical strings, they would likely not accept the foil strings.

## Materials

The experiment was programmed and run in Praat version 5.4 (Boersma & Weenink, 2014) and was conducted on a desktop computer running Windows 7, with a 17-inch monitor (refresh rate: 60 Hz; resolution: 1280 x 1024 pixels).

## Procedure

Participants sat approximately 50 cm from the screen in a quiet experimental room and listened to the stimuli presented through headphones (Sennheiser HD 201). Participants responded by clicking boxes indicating their responses on the computer screen via mouse.

Participants were first familiarized with the Fibonacci-grammatical sequence then tested in one of the two testing paradigms, familiarized again, and finally tested with the other testing paradigm. The procedure of the experiments is explained via the exact instructions given to participants.

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.*

Participants saw different instructions before each of the two test phases.

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 one-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?*

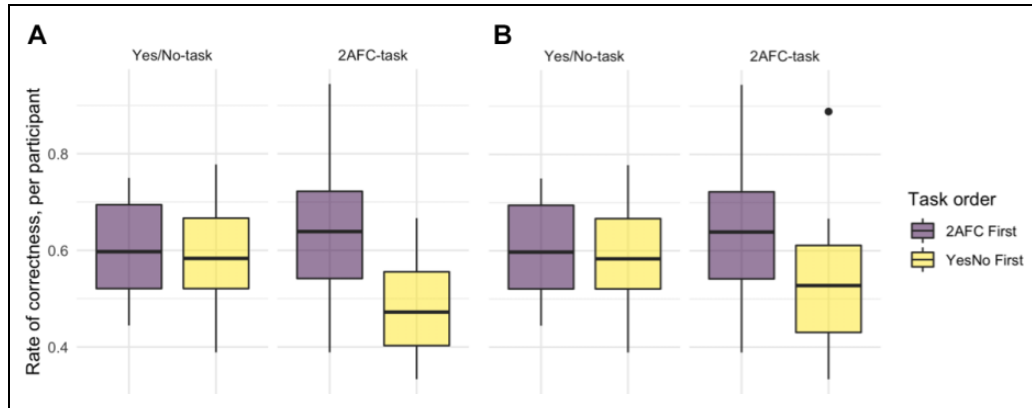
Before the Yes/No test phase, participants saw the following instructions: *The test phase will now begin. You will hear 36 test sounds. For every sound, listen carefully and indicate whether it follows the same rhythm as during the listening phase by clicking "yes" or "no".*

In both testing orders, the response screen included two boxes: either two boxes showing the words "First" and "Second" for the 2AFC task, or two boxes showing the words "Yes" and "No" for the Yes/No task. At the bottom of the screen, participants saw a Likert scale (Likert, 1932) for rating the sureness of each response with the following instruction: *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.*

## Analysis of Data

Participants' responses in each case were recorded and our output variable per trial was correctness of response. For statistical analysis, we summarized the categorical response per participant and per task, resulting in *rate of correctness* as our dependent variable. Rate of correctness was computed as the sum of correct trials per participant and task divided by number of observations per participant and task. For the one-sample *t*-test of all trials, independent of test, we computed the rate of correctness per participant, but not per task. Participants' sureness scores were also analyzed. Rate of correct responses was analyzed using one-sample *t*-tests. A mixed ANOVA was performed to assess the effect of our factors, namely task, task order, and musical training, on participants' performance.

Below, we report on all experimental measures collected. Note that Praat software, which we used for running the experiment, automatically collects reaction times, but these were not part of our planned comparisons, so we do not report on them. We also report on all relevant participant questionnaire responses for which we had clear hypotheses and planned comparisons, including namely musical background. Participant questionnaire responses on dyslexia and hearing problems were used for exclusion. Finally, other questionnaire responses, including handedness, language background, and study background, were not analyzed, but could be useful for future exploratory analysis and are included in our data file, available at



**Figure 2.** Rate of correct responses, across task and task orders (A: excluding outliers, B: including outliers). Outliers are denoted by points (in panel B of this figure at the same position), medians are reflected by the horizontal line, and boxes represent the interquartile range.

[https://osf.io/s2f3h/?view\\_only=3191f5635f4b4dc48093ac36950f733f](https://osf.io/s2f3h/?view_only=3191f5635f4b4dc48093ac36950f733f), along with the questionnaire in its entirety.

The data was analyzed using R version 3.6.2 (R Core Team, 2019) and R packages: ggpubr (Kassambara, 2019a), rstatix (Kassambara, 2019b), and tidyverse (Wickham et al., 2019).

## Results

Our data were divided along three dimensions (factors). The first two independent variables were task (which task was being performed at each trial: 2AFC-task or Yes/No-task), and task order (which task was performed first: 2AFC-first or Yes/No-first). Furthermore, participants were categorized by whether they have previously received musical training or not.

### Rate of Correctness

To test our main hypothesis of participants being able to discriminate between the Fibonacci-grammatical and pseudo-Fib foil sequences, we compared each participant's overall rate of correctness and the rate of correctness both in the Yes/No-task and 2AFC-task to chance level rate of correctness (null hypothesis) using one-sample *t*-tests. We decided to compare our test means against an expected population mean of 0.5, as each task had two options and we would therefore expect a rate of correctness of 0.5 if participants were unable to discriminate the Fibonacci-grammatical sequences from the foil sequences. In Figure 2B, all rates of correctness including outliers are displayed per participant and per task, with task order reflected by different colors. The rate of correctness for both tasks and overall were normally distributed, as assessed by a Shapiro-Wilk test (all  $p > .05$ ). Overall, the rate of correctness (per participant and task) was indeed statistically significant above chance level (0.5),  $t(31) = 5.6605$ ,  $p < .001$ , Cohen's  $d = 1.00$ . Both in the Yes/No-task and 2AFC-task, the rate of correctness was also

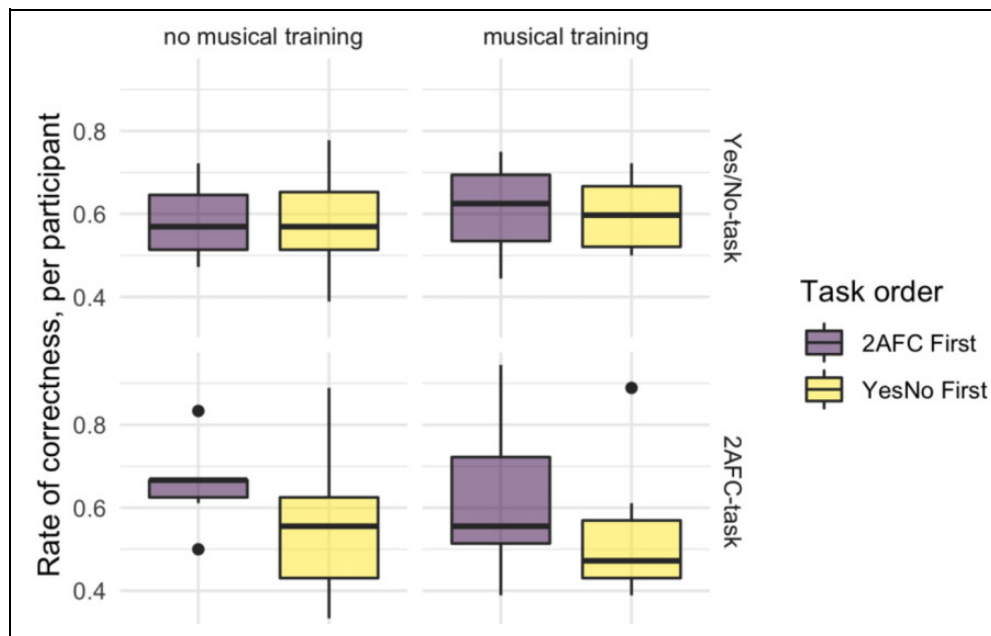
statistically significant above chance level for the Yes/No task and for the 2AFC task as compared to an expected mean rate of correctness of 0.5,  $t(31) = 5.4415$ ,  $p < .001$ , Cohen's  $d = 0.96$  (YN-task) and  $t(31) = 3.0528$ ,  $p < .005$ , Cohen's  $d = 0.54$  (2AFC-task).

The observations of Yes/No-task and 2AFC-task are not completely independent when considering overall rate of correctness and rate of correctness in only one task, as each participant took part in both tasks. This implies that the results for overall rate of correctness need to be considered more carefully, as the one-sample *t*-test assumes independence of the data points. However, considering the fact that participants performed significantly above chance performance in both tasks separately, we assume that also the overall performance is indeed above chance performance.

### Task and Task Order

We were further interested in testing how task and task order affect participants' ability to discriminate between Fibonacci-grammatical and pseudo-Fib foil sequences. We did not have any prior expectations of the rate of correctness being higher in one task over another, but we did expect to find a learning effect from the first to the second task within participants. We compared the rate of correctness across participants between tasks (within-participant factor) and task order (between-participant factor). There were two outliers (removed by at least 1.5\*interquartile range from the interquartile range), both in the Yes/No-first order and 2AFC task (see Figure 2). As we were sure the outliers could not have been caused by technical issues, and we did not set the 1.5 threshold before running the experiments, we were skeptical about interpreting these values as outliers, and therefore report our analyses both with and without the outliers included.

We performed a two-way mixed ANOVA to examine the effect of task and of order of task on rate of correctness per participant and task. When we included the outliers in



**Figure 3.** Rate of correct responses, across musical training (horizontal), task order (colors) and task (vertical). Outliers are denoted by points, medians are reflected by the horizontal line, and boxes represent the interquartile range.

the analysis, there was no statistically significant effect of either task [ $F(1,30) = 0.082, p > .05$ ] or task order [ $F(1,30) = 2.164, p > .05$ ], nor an interaction effect of the two factors [ $F(1,30) = 1.594, p > .05$ ] on the rate of correctness. However, when we did exclude the previously mentioned outliers, we found a statistically significant effect of task order on the rate of correctness [ $F(1,28) = 10.713, p < .005$ ], while the effect of task [ $F(1,28) = 0.505, p > .05$ ] and the two-way interaction effect [ $F(1,28) = 2.776, p > .05$ ] remained statistically non-significant. Figure 2 reflects how the removal of the outliers shifts the median of the data, potentially causing the observed main effect of task. This becomes especially obvious in panel B, where the outliers, which are both in the same position, shift the medians of the rate of correctness towards the medians observed across tasks and task orders. Due to this substantial difference in results between the dataset including the two outliers compared to the dataset excluding the two outliers, we cannot confidently accept or reject the null hypothesis of no main effect of task. While the pattern of a lower rate of correctness in the Yes/No-first order that is observed when we exclude the outliers can also be observed as a general tendency in the data including the outliers (see Figure 2), the two outliers cannot be theoretically dismissed as data points.

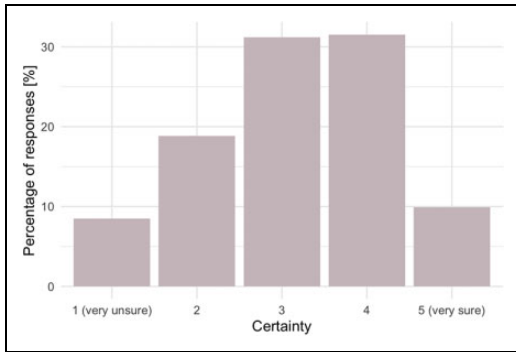
### Musical Training

Apart from task and task order as independent variables, we were also interested in the effect of musical training as a between-subjects factor on observed rate of correctness. However, when summarizing data across task\*task order\* musical training, this resulted in eight combinations of

factor levels. This led to a smaller number of observations per combination of factor levels, and one of the combinations (2AFC first, 2AFC task, no musical training) had a non-normal distribution (see Figure 3). For this reason, we decided to carry out our main analysis only on the two factors task and task order (see previous paragraph), which we were more strongly interested in. However, as we also predicted previous musical training to influence rate of correctness positively compared to no previous musical training, we decided to also perform a three-way mixed ANOVA with task as a within-subject factor and task order and musical training as between-subjects factor. As this analysis includes one non-normal distribution, we explicitly only interpret these results tentatively. Figure 3 shows the rate of correctness across musical training, task, and task order. An analysis running all data points, including outliers, did not show a statistically significant effect of either musical training [ $F(1,28) = 0.019, p > .05$ ], task [ $F(1,28) = 0.018, p > .05$ ], or task order [ $F(1,28) = 2.004, p > .05$ ] on rate of correctness, nor any interaction effects (all  $p > .05$ ). Removing outliers resulted in too few data points ( $n = 3$ ) for the no musical training-2AFC first-2AFC task group to carry out the analysis. This indicates that our expected advantage of participants with musical training was not reflected in the data. However, as discussed, assumptions of the mixed ANOVA were not met, and therefore we can only interpret these results tentatively.

### Participant Response Certainty

To give an indication of how certain participants were about their decision per trial, we analyzed the percentage



**Figure 4.** Percentage of responses per certainty category, across all trials.

of the sureness that participants had indicated per response, as can be seen in Figure 4. Overall, participants felt “5 – very sure” about 9.9% of their responses, “4 – somewhat sure” about 31.5% of their responses, “3 – not sure” (i.e., neutral) about 31.2% of their responses, “2 – somewhat unsure” about 18.9% of their responses, and “1 – very unsure” about 8.5% of their responses. If participants were neither especially sure or unsure, we would expect a mean certainty response of 3. However, a one-sample Wilcoxon signed-rank test of overall certainty revealed certainty to be statistically significantly above the expected mean of 3 [ $p < .001$ , effect size  $r = 0.129$ ], indicating that participants had a tendency to be more certain than they were uncertain about their choices. Their certainty scores support the finding that participants were sensitive to the difference between the Fibonacci-grammatical sequences and the pseudo-Fib foil sequences. However, it must be noted that effect size is small, meaning that participants were not very confident in their choice.

## Discussion

In our experiment, we tested participants’ capacity to discriminate between a series of Fibonacci-grammatical drumming sequences and pseudo-Fib foil sequences not part of the Fibonacci-grammatical space but sharing surface properties with the grammatical stimuli. We found that participants were sensitive to the familiarization Fibonacci grammar, and were overall capable of distinguishing the grammatical and ungrammatical test stimuli. A one-sample  $t$ -test of rate of correctness per participant compared to an expected outcome of 0.5 as the null hypothesis 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 (each approximately 0.6).

While the recognition rate we find is significant and systematic, it is not at the level shown in Shirley (2014) where participants had a mean identification accuracy of Fibonacci-grammatical strings in the range of 70–80% in a 2AFC task (Shirley, 2014, Chapter 3). These outcomes

suggest that participants picked up some of the grammatical regularities they were exposed to, although not robustly. This discrepancy points to a potential facilitating role of speech stimuli for structure learning and recursive processing. Nevertheless, the fact that our participants show more sensitivity to the Fibonacci grammars than would be predicted by chance when they are instantiated with drum sound stimuli indicates that L-systems as a class have the potential to be a useful tool both for artificial grammar learning experiments and musical processing experiments, and can complement simpler grammars across domains.

In addition, we did not find any effect of task, nor an interaction between task and task order. However, when excluding outliers, there was a statistically significant effect of task order, but we did not find this effect when analyzing all data. We therefore remain agnostic about the effect of task order. Furthermore, while we set out to investigate whether musicianship would affect ability to process recursion in musical stimuli, it proved to be difficult to analyze the effect of musical training on rate of correctness, as some variables were not normally distributed, potentially due to a relatively small number of observations when choosing a larger number of factors for the ANOVA. An analysis including outliers showed no main effects, nor any interaction effects. This would support our hypothesis of no effect of task, but discredit our hypotheses of a positive effect of musical training and an effect of learning from the first to the second task, the latter of which would have been reflected in an interaction effect between task and task order. ¶ Several potential reasons may explain the lack of clear effects of any of our hypothesized predictor variables. First, our study may be underpowered. Considering the complexity of the grammar, a larger sample size could have helped in detecting a small effect. Second, learning may have been occurring over trials within one or both tasks; a larger sample size would have allowed for finer-grained trial-by-trial analyses, showing potential effects of learning. Finally, there may have been a high interindividual variability: individual differences are common sources of variance in grammar learning experiments across species, and they may obscure effects of task and stimuli (Danner et al., 2017; Kepinska, et al., 2017; Ravignani et al., 2015).

Keeping these disclaimers and caveats in mind, there was evidence of sensitivity of participants to the target Fibonacci-grammatical strings overall. The grammar is too complex for participants to be able to explicitly state what rule generates it. 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 strings presented to the participants were potentially part of the Fibonacci-grammatical space, they were unable to discriminate them from the exposure grammar. However, in the present work, where we assumed the pseudo-Fib foil strings were not a part of the Fibonacci-grammatical space but still

shared surface properties with the Fibonacci-grammatical strings, participants were able to pick out this grammar as different and ungrammatical, in both a Yes/No task and a 2AFC task.

One possible alternative explanation for participants' success may be a result of a sensitivity to the repetition of 15-element-long sequences in the foil strings that was not present in the Fibonacci-grammatical strings.<sup>2</sup> However, based on previous literature related to the limits on serial recall, it is unlikely that listeners would be able to hold such long sequences in memory to notice the repetition spanning 15 elements. The longest sequences that are held in memory have been found to be of seven items (e.g., Aaronson et al., 1971). Serial recall of longer sequences has found to be aided by chunking into adjacent pairs (with four pairings of two items estimated to be the maximum capacity of young adult participants who give their full attention; Naveh-Benjamin et al., 2007). Even if participants were able to process the stimuli in this way, they would be confronted with the differences in chunks between the Fibonacci-grammatical and the foil stimuli when performing this implicit calculation. Therefore, in order to discriminate longer items, some sort of understanding of the structure of the sequences should therefore be implicitly learned. This would strengthen the argument for a sensitivity to the rules of the Fibonacci sequences, even if this sensitivity takes the form of understanding that there are some rules at play in one type of stimulus, but not in the other. While we leave the door open for the possibility that the repetition present in the foil strings may have contributed to or may explain the success of participants, it is not likely. Such a hypothesis could be directly tested by varying the lengths of sequences of the foil sequences to see the impact on discrimination abilities.

While we did not find that task on its own had an effect on learning, there was potential evidence of better outcomes when participants were performing the 2AFC task as the first task (see Figure 2). This may be because although both tasks are considered recognition tasks, the Yes/No task may be characterized as a categorization 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 the correct one must then be identified. As such, they tap into different recognition mechanisms. The 2AFC task performed immediately after exposure may therefore be better suited to tap into sensitivities that participants may have gained during the exposure phase (Jang et al., 2009).

In order to better understand how processing of recursion unfolds in real time and to better catch nuances in performance, improvements to these experiments can be made. To this end, tasks that incorporate immediate reaction times, such as electrophysiological recordings or serial reaction time tasks, should be employed. We have already started working in this direction with a simultaneous

rhythmic tapping experiment (Minnema et al., 2018). Such online measures should give us a better understanding of how participants process complex grammars and rhythmic sequences.

Another detail that should be addressed in future work is the number of items in the test strings, which in the present work is 50. This is not a Fibonacci number and therefore, the grammatical test items could be considered substrings of a grammatical string, strictly speaking.<sup>3</sup> This is comparable to the common practice of fade-in and fade-out that is used in auditorily presented sequence learning and artificial grammar learning tasks, in which the participant hears a substring of the grammatical items. While this detail is unlikely to have had an effect on the participants' ability to perform our task, it is formally important. Future work would be better served to push the foil sequences to be closer in number of elements to a complete Fibonacci-grammatical string.

Finally, while we are not mathematically certain that our foil test strings are not also a part of the Fibonacci grammatical space, we do make the assumption that they are not (see stimuli section). On the other hand, if we are wrong and the foil is also a part of the Fibonacci-grammatical space, our participants would be showing evidence of discrimination between two different Fibonacci grammars, which would also be novel and have further theoretical implications. While we cannot make this claim due to the lack of mathematical proof available, we can nevertheless conclude that participants were able to show evidence of sensitivity to recursive properties found in a set of Fibonacci-grammatical strings in the musical domain.

### Acknowledgements

We thank Doug Saddy, Liz Shirley, and Diego Gabriel Krivochen for valuable discussion on the experiments.

### Contributorship

Andreea Geambaşu, Andrea Ravignani, and Clara C. Levelt conceived the experiments and design. Andreea Geambaşu programmed the experiments. Laura Toron ran the experiments and analyzed the data. Andrea Ravignani created the stimuli. All authors contributed to and approved the writing and revision of the manuscript.

### Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.



### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Andreea Geambaşu and Clara C. Levelt were supported by the



NWO Vrije Competitie grant 360.70.452 to Clara C. Levelt. Andrea Ravnani was supported by ERC grants 283435 ABA-CUS (to Bart de Boer), 230604 SOMACCA (to W. Tecumseh Fitch) and ESF grant 5544 INFY (to Andrea Ravnani), and funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665501 with the research Foundation Flanders (FWO) (Pegasus2 Marie Curie fellowship 12N5517 N).

### ORCID iDs

Andreea Geambaşu  <https://orcid.org/0000-0002-5883-9875>  
 Laura Toron  <https://orcid.org/0000-0003-2566-7699>

### Action Editor

Samuel Mehr, Harvard University, Department of Psychology.

### Peer Review

Fabian Moss, École Polytechnique Fédérale de Lausanne, Digital Humanities Institute.

One anonymous reviewer.

### Notes

1. Given an alphabet of symbols and any string of those symbols, a finite automaton is a deterministic finite state machine which accepts or rejects a string based on the sequence and adjacency relation of the symbols. Here deterministic means that if the automaton accepts a string once, it will always accept it. For instance, a finite automaton with only one state and only one transition—for example, “a”—will accept all and only strings of any length composed of all “a” symbols.
2. We thank an anonymous reviewer for this suggestion.
3. We thank an anonymous reviewer for this suggestion.

### References

- Aaronson, D., Markowitz, N., & Shapiro, H. (1971). Perception and immediate recall of normal and “compressed” auditory sequences. *Perception & Psychophysics*, *9*(4), 338–344.
- Bahlmann, J., Gunter, T., & Friederici, A. (2006). Hierarchical and linear sequence processing: An electrophysiological exploration of two different grammar types. *Journal of Cognitive Neuroscience*, *18*(11), 1829–1842.
- Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program]. Version 5.4. <http://www.praat.org/>
- Danner, D., Hagemann, D., & Funke, J. (2017). Measuring individual differences in implicit learning with artificial grammar learning tasks. *Zeitschrift für Psychologie*, *225*, 5–19.
- Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., & Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children, US adults, and native Amazonians. *Science Advances*, *6*(26), eaaz1002.
- Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*(3), 329–364.
- Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1933–1955.
- Geambaşu, A., Ravnani, 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), 1–7. <http://doi.org/10.3389/fnins.2016.00281>
- Gollin, E. (2008). Near-maximally-distributed cycles and an instance of transformational recursion in Bartók's Etude Op. 18, No.1. *Music Theory Spectrum*, *30*(1), 139–151.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.
- Hochmann, J. R., Azadpour, M., & Mehler, J. (2008). Do humans really learn  $A^nB^n$  artificial grammars from exemplars? *Cognitive Science*, *32*(6), 1021–1036.
- Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291–306. <http://doi.org/10.1037/a0015525>
- Kassambara, A. (2019a). ggpubr: “ggplot2” based publication ready plots. R package version 0.2.4. <https://CRAN.R-project.org/package=ggpubr>
- Kassambara, A. (2019b). rstatix: Pipe-friendly framework for basic statistical tests. R package version 0.3.1. <https://CRAN.R-project.org/package=rstatix>
- Katz, J., & Pesetsky, D. (2009). The recursive syntax and prosody of tonal music. In *Recursion: Structural complexity in language and cognition* conference, University of Massachusetts (Amherst).
- Kepinska, O., de Rover, M., Caspers, J., & Schiller, N. O. (2017). On neural correlates of individual differences in novel grammar learning: An fMRI study. *Neuropsychologia*, *98*, 156–168.
- Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443–15448.
- Krivochen, D., & Saddy, D. (2018). Towards a classification of Lindenmayer systems. *arXiv preprint arXiv:1809.10542*.
- Lai, J., & Poletiek, F. H. (2013). How “small” is “starting small” for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology*, *25*(4), 423–435.
- Lerdahl, F., & Jackendoff, R. S. (1996). *A generative theory of tonal music*. MIT Press.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 1–55.
- Lindenmayer, A. (1968). Mathematical models for cellular interactions in development I: Filaments with one-sided inputs. *Journal of Theoretical Biology*, *18*(3), 280–299.
- Losada, C. C. (2007). K-nets and hierarchical structural recursion: Further considerations. *Music Theory Online*, *13*(3).
- Loy, G., & Abbott, C. (1985). Programming languages for computer music synthesis, performance, and composition. *ACM Computing Surveys (CSUR)*, *17*(2), 235–265.
- Manaris, B., & Brown, A. R. (2014). *Making music with computers: Creative programming in Python*. CRC Press.



