Structural learning in Lindenmayer grammars

Schmid, Samuel

This publication URL: https://archive-ouverte.unige.ch/unige:171340
Publication DOI: 10.13097/archive-ouverte/unige:171340

UNIVERSITÉ DE GENÈVE
FACULTÉ DE PSYCHOLOGIE ET DES SCIENCES DE L'ÉDUCATION

Section de Psychologie

Sous la direction du Dr Julie Franck

STRUCTURAL LEARNING IN LINDENMAYER GRAMMARS

THESE

Présentée à la
Faculté de psychologie et des sciences de l'éducation
de l'Université de Genève
pour obtenir le grade de Docteur en **Psychologie**

par

Samuel SCHMID

de

Suisse (Bern)

Thèse No 851

GENEVE

Août 2023

N° étudiant 13-338-330

Acknowledgments

I would like to express my gratitude to the committee and jury members, Daphnée Bavelier, Pierre Barouillet, Doug Saddy and Julie Franck, for their valuable time and effort in reading and evaluating this work. I am especially grateful to Julie Franck and Doug Saddy for guiding me through this project. Without their commitment and support, this work would not have been possible.

I would also like to thank the members of the laboratory of psycholinguistic of the University of Geneva. In particular, I thank my friend and colleague David Gyorgy, with whom I shared the joys and frustrations of scientific research as well as a significant amount of alcoholic beverages. I would like to express my gratitude to Diego Krivochen for engaging in insightful conversations and providing valuable comments on certain chapters of this thesis. Additionally, I extend my thanks to Denis Delfitto, Maria Vender, and Beth Phillips for their valuable discussions and constructive feedback.

Finally, I would like to thank from the bottom of my heart my friends and family who have always supported me and given me the strength to continue in moments of doubt. All my thoughts go to my beloved wife, Lydie, and to our wonderful little monster, Oscar.

Ce travail est dédié aux trois personnes les plus merveillleuses de la terre, Lydie, Oscar et Biscotte.

Abstract

This thesis investigates whether and how the human cognitive system to extract recursive nested structures from a highly simplified input where hierarchy is marked only by sequential order information. Elaboration and processing of hierarchical representations is involved in many domains of human cognition (Martins, 2012; Uddén et al., 2020), the most notable being language. Natural languages are characterized by structural dependencies in which constituents are linked to each other in such a way that sentences cannot be reduced to the linear relationships between these constituents. Therefore, to correctly interpret a sentence, the cognitive system cannot rely solely on the linear relationship between words but must go beyond this linearity and extract the underlying hierarchical structure of the sentence (Chomsky, 1957). However, demonstrating the building of hierarchical representations in sequences' processing has proven difficult (Levelt, 2020). The difficulty stems from the complexity of implementing hierarchical structure in artificial settings as well as from methodological issues associated with the conventional habituation/discrimination testing procedure. As a result, effects that have been attributed to hierarchical learning can also be attributed to the encoding of the surface properties of the input (Perruchet, 2005).

Our question is thus whether the cognitive system develops hierarchical representations when processing non-linguistic sequences. To address this question, we investigated the processing of aperiodic and self-similar binary strings generated by the Fibonacci grammar (Lindenmayer, 1968). Instead of the habituation/discrimination paradigm, we evaluated the extraction of hierarchical structures by incorporating the strings generated by this grammar into a serial reaction time (SRT) task. By leveraging the properties of the Fibonacci grammar and the SRT task, we were able to investigate the elaboration of hierarchical representations while controlling for the use of surface strategies.

Three main questions were investigated concerning the processing of the Fibonacci grammar. We first examined whether participants developed a hierarchical representation of Fib-generated strings during an SRT task. Results revealed that participants' pattern of anticipation could not be explained solely by "flat" statistical learning processes. Instead, anticipation appeared to be based on hierarchical assumptions. Additionally, we observed that participants exhibited sensitivity to the grammar's constituent structure, suggesting that they organized the input into embedded constituents. Secondly, we explored the extent to which one of the formal properties of the Fibonacci grammar, namely the isomorphism between derivational order and sequential order, plays a role in hierarchical structure extraction. In two experiments, we compared participants' processing of the Fibonacci grammar with that of an alternative grammar, the Skip grammar, which does not exhibit this isomorphism. Results showed that participants extracted a hierarchical structure during the processing of both grammars, suggesting that isomorphism is not a key property for hierarchical structure extraction. Finally, we explored the impact of presentation rate on the extraction of hierarchical structure by manipulating the duration of the Response-to-Stimulus Interval (RSI). Multiple hypotheses have been put forward in the literature to account for the influence of RSI duration on sequence learning in the SRT task (Frensch & Miner, 1994; Willingham et al., 1997). However, this question has never been addressed from the perspective of hierarchical structure extraction. We found that RSI duration affected hierarchical elaboration in a non-linear way, with participants building higher hierarchical structures with an RSI of 250 ms compared to RSIs of 1000 ms and 100 ms. This finding suggests the presence of an optimal temporal window for sequence learning in the SRT task. This U-shaped effect cannot be accounted for by any of the existing hypotheses on the influence of RSI duration on sequence learning in the SRT task. We hypothesized that this effect results from the tension between the cognitive system's limited encoding capacity and the amount of information per unit of time delivered to the system.

List of abbreviations

Fib grammar : Fibonacci grammar

L-systems : Lindemayer systems

MS : Milliseconds

RSI : Response-to-Stimulus Interval

RT : Reaction Time

SOC : Second Order Conditional

SRT : Serial Reaction Time

Statement of contributions

Chapters 2, 3, and 4 of this thesis were conceived and authored as a separate research article. The principal investigator of each article is Samuel Schmid. Chapter 2 is a reformatted version of the published article :

Schmid, S., Saddy, D., & Franck, J. (2023a). Finding Hierarchical Structure in Binary Sequences : Evidence from Lindenmayer Grammar Learning. *Cognitive Science*, *47*(1), e13242. https://doi.org/10.1111/cogs.13242

Chapter 3 is a reformatted version of the article submited on January 23, 2023 to PLOS ONE under the title: "Uncovering hierarchical structure through statistical dependency". The article is currently undergoing revision but is accessible as a preprint :

Schmid, S., Saddy, D., & Franck, J. (2023b). *Uncovering hierarchical structure through statistical dependency*. PsyArXiv [Preprint]. https://doi.org/10.31234/osf.io/mre4h

Chapter 4 of this thesis is planned for submission as a research article to an academic journal in the near future.

CONTENTS

Chapter 1. General introduction

1.1 Introduction

To predict the course of a sequence of events over time, the cognitive system must infer from the surface properties of the signal its underlying organization, i.e., make an approximation of the underlying structure of the signal and use this mental model to generate predictions allowing a better adaptation. The difficulty is that this underlying structure is not directly accessible in the signal itself but must be inferred from it. Consider the situation where the cognitive system receives as input the following speech signal: *the picture of the neighbor is gone*. Any individual with a sufficient command of the English language will be able to correctly answer the question "*is the picture gone?*" despite the fact that the signal contains the sequence "*the neighbor is gone*". Although such behavior may seem trivial at first glance, the computational problem that the cognitive system faces in such a situation is quite complex. The input the system receives is fundamentally linear : words unfold in time. However, to understand that what is gone is the picture of the neighbor and not the neighbor himself, the system cannot rely on the adjacent relationship between words. This is because the syntactic relationship between words is not proportional to the distance between them, it is not because two words are temporally close that they are syntactically related. Thus, in order to correctly interpret this sentence, the system has to extract its underlying hierarchical structure.

Most theories of language processing assume that human language production and comprehension cannot be captured by the mere concatenation of consecutive items but must be mentally represented as recursive nested structures (Chomsky, 1957; Lashley, 1951; Simon, 1962). However, demonstrating the extraction of nested structures in sequence processing through experimental methods has proven challenging. To investigate the mechanisms involved in hierarchical structure extraction, researchers have resorted to creating simpler artificial systems that allow for highly controlled settings. However, there is a limited number of empirical studies in the field of artificial grammar learning (AGL) that provide conclusive evidence of nested structure extraction (Fitch, 2014; Honing & Zuidema, 2014; Kovács & Endress, 2014; Levelt, 2020). The reason for this situation comes from the fact that in the test case classically used with this paradigm, a sequence can be processed without requiring the construction of a nested structure. Instead, alternative and potentially simpler ways of representing the sequence can lead to comparable learning performance (Dehaene et al., 2015; Maheu et al., 2019; Uddén et al., 2020).

The objective of this thesis is to explore the ability of the human cognitive system to extract hierarchical structure from a sequentially presented input. Our experimental research focuses on the underlying cognitive mechanisms involved in this process. It should be noted that we make no assumptions about an equivalence between the cognitive mechanisms used by participants during our experimental investigations and those involved in language processing. We explored the extraction of hierarchical structure in sequences generated by a grammar belonging to the Lindenmayer formalism (Lindenmayer, 1968) : the Fibonnacci Grammar. Sequences generated by this grammar are aperiodic and self-similar. Instead of the habituation/discrimination paradigm classically used, we implemented the sequence generated by this grammar in a Serial Reaction Time task (SRT) (Nissen & Bullemer, 1987). By leveraging the features of the Fibonacci grammar and the SRT task, we were able to assess the extraction of hierarchical structures without facing the challenges commonly encountered in the existing literature on this topic.

The subsequent sections of this introduction will address several key points. Firstly, we will clarify the distinction between hierarchical and algebraic representations. Following that, we will briefly explain why the AGL paradigm is not a suitable approach for investigating hierarchical representation. Then, we will describe the methodology retained in this work and how it overcomes the limitations of the AGL paradigm. Finally, we will outline the main research questions addressed in the following chapters and provide a preview of the key findings.

1.2 Hierarchical structure in sequence processing

Demonstrating the learning of hierarchical structure in sequence processing has proven challenging. The main difficulty is conceptual. Just as the cognitive system does not have direct access to the underlying structure of the signal, the researcher in cognitive science does not have direct access to the mental representations of the individuals he studies. As a result, the architecture of mental representations must be deduced from observable behaviour. The problem is that the output of a system is often compatible with several generators. For example, humans are relatively good at estimating where a ball will fall after being thrown. However, this does not imply that Newtonian laws of physics are encoded somewhere in the brain as a generator that would work on the basis of simpler heuristics could account for this behavior. The same situation arises in the study of sequence processing: the same sequence can be underpinned by several representational architectures (Uddén et al., 2020).

1.2.1 Defining hierarchical structure

The term "hierarchical structure" has occasionally been confounded with that of "algebraic pattern" (Koch & Hoffmann, 2000). It is therefore important to define what we mean hierarchical structure in the context of the present work. In this regard, Dehaene et al. (2015) presented a taxonomy that categorizes various types of internal representations that can be generated from a sequence. As per the taxonomy proposed by these authors, the differentiation between algebraic patterns and hierarchical structures can be outlined as follows. An algebraic pattern representation is a type of representation that encodes the relationships between the elements of a sequence independently of their perceptual expression; in other words, the relationships are encoded as variables whose arguments can take several values. Algebraic representations allow to explain the generalization of a structure to new material. An example of an algebraic representation developed in sequence processing is the learning of the AAB structure. Marcus et al. (1999) investigated seven-month-old children's sensitivity to the structure of trisyllabic pseudowords, exposing them to pseudowords with the AAB structure (such as "*duduba*" or "*pipiro*"), where the first two syllables were identical and the third was different. The results showed that when presented with new pseudowords made of new syllables, the children could distinguish between pseudowords that followed the AAB structure and those that followed the ABA structure. Insofar as test strings were composed solely of new syllables, the authors argued that children had developed an algebraic rule-like representation of the AAB pattern. However, the representation of the AAB structure, even though abstract, cannot be called hierarchical because the relationship between the categories is analogous to the order of presentation of the sequence elements. This is where a hierarchical representation differs from an algebraic representation. A hierarchical representation (also referred to as *nested* representation) is a type of representation where elements can be nested into a higher order constituent. This permits the encoding of a relationship between two elements regardless of the distance between them. Thus, arbitrarily large material can intervene between the linked elements without altering the dependency relationship. For example, complex structural dependencies that characterize natural languages, like the subject-verb agreement dependency cannot be captured by an algebraic pattern. This is due to the fact that the elements involved in agreement are constituents that may encompass nested constituents at lower levels, allowing for arbitrarily large amounts of material to occur between the subject and the verb without disrupting the dependency relationship.

1.2.2 Limitations of the AGL paradigm

One of the widely used paradigms to study the learning of hierarchical structure is the AGL task. In a typical AGL experiment, participants are first presented with a set of training strings that are generated by a specific grammar. They are then asked to use the acquired knowledge to recognize new strings that are either generated by the same grammar (grammatical strings) or generated by a different grammar (ungrammatical strings). The task is typically performed without explicit instruction or feedback, and participants rely on their implicit knowledge to perform it. In a seminal study using this paradigm, Reber (1967) found that after the exposure phase, participants were able to distinguish new grammatical strings from ungrammatical strings. Successful discrimination between grammatical and ungrammatical strings has been replicated in many AGL studies, however, the type of representations that are developed in this task as well as the learning mechanism that underlies the development of these representations remains a matter of debate (Gervain et al., 2020; E. M. Pothos, 2007; Schiff & Katan, 2014; ten Cate et al., 2020; Trotter et al., 2020, see Chapter 2 for a detailed description of the debate).

Five main reasons may explain why the AGL paradigm failed to provide unequivocal evidence in favour of hierarchical learning. The first is the difficulty of ensuring that the discrimination between grammatical and ungrammatical test strings is possible only if the underlying structure of the training material has been extracted by the participants. Indeed, if ungrammatical test strings exhibit distinct surface properties that are not present in grammatical strings, then their rejection may be attributed to the representation of these surface properties rather than their lack of grammaticality. This necessitates that the dimension along which grammatical and ungrammatical test strings diverge is not perceptually expressed in the input signal. However, since a hierarchical structure is in the mind of the one that processes the signal and not in the signal itself, the grammaticality of a test string is necessarily judged on the basis of whether or not its surface properties (i.e., what the parser has access to) allow the extraction of a structure similar to the target structure. Thus, a foil can only be judged as such if its surface properties do not allow the extraction of a grammatical structure, implying that a foil that does not diverge in its surface properties with a grammatical string is in fact, a grammatical string. Modifying the perceptual attributes of stimuli between the exposure and test phases, such as using new syllables or modifying the shape or color, does not offer a solution to this problem. This is because at least one surface property of the training material must be retained in the test materials in order to discriminate between grammatical and ungrammatical test strings. For example, in the Marcus et al. (1999) study, the AAB pseudowords used in the exposure and test phases, despite consisting of different syllables, share a surface property: the order in which the syllables are presented.

The second limitation of the AGL paradigm is that the mechanisms used during the exposure phase may differ from those used in the test phase. This task asks participants to compare the representation developed during the exposure phase with the test string and to judge, on the basis of their degree of similarity, whether or not they belong to the same category. The assumption is that the same structure extraction mechanisms are used in both phases. However, only the mechanisms involved in the test phase are measured, and the most efficient strategy for correct discrimination may not necessarily involve extracting the underlying structure of the test string. As a result, even if participants have successfully extracted the underlying structure from the training materials, they may rely on different strategies to judge the grammaticality of the test strings. Just as it is impossible to determine the exact properties that lead to the rejection of an ungrammatical string (i.e., whether based on its higher-order or surface properties), it is impossible to determine on the basis of which properties a grammatical string is judged as grammatical: this may be on the basis that the grammatical string shares surface properties with previously learned materials, or on the basis that an identical hierarchical structure can be extracted, or on both.

The third limitation is that the test materials may contaminate the knowledge developed during exposure as participants continue to learn. As a result, if too many test items are administered, the contamination effect may become too prominent, leading to an inaccurate assessment of learning. The number of test strings that can be presented is therefore limited, which has a direct impact on the statistical power of the dataset, as a smaller number of test items can reduce the reliability and validity of the results.

The fourth limitation of the AGL task is that learning is measured offline, after exposure, which makes it challenging to assess the evolution of learning over time using this paradigm. Incorporating alternating learning and testing phases may be a solution, but it can introduce confounding factors like contamination from test strings, making data interpretation difficult. Another approach could be to manipulate the duration of exposure across participants, ensuring only one test phase per length of exposure. However, despite its theoretical feasibility, practical implementation of this design may pose challenges as it would require different conditions for each time point in the learning trajectory being measured and to our knowledge, no study has used such a design.

Finally, the quantification of learning, i.e., the extent to which participants have learned the target structure, is often overlooked in the AGL field (Franck et al., 2016). This oversight stems from a vision of learning as a discrete process, despite being measured continuously. Typically, successful learning is defined based on above-chance performance, resulting in a binary view where participants are deemed to have either learned or not learned the target structure. Therefore, if a participant's performance is 58% above chance, they may be categorized as having learned the target structure, while another participant with a performance of 80% above chance also fall in the same category. This approach oversimplifies the complexity of the learning process and ignore the potential differences in the quality and depth of learning.

1.2.3 Addressing the limitations of the AGL paradigm

The AGL paradigm's challenge lies in its two-stage organization, consisting of a learning phase followed by a testing phase. If learning could be measured continuously directly during the exposure phase, the aforementioned limitations could be circumvented. First, direct measurement of learning eliminates the need to present ungrammatical strings to participants. Second, it is not necessary to assume that the mechanisms at play during exposure are identical to those involved in the test phase since the test phase is eliminated. Third, contamination of learned representations by test strings during the testing phase is avoided, again, because the test phase is no longer present. Finally, the continuous measurement makes it possible to finely estimate the evolution of learning over time.

In order to measure learning directly during exposure, two key components are required. First, a sequence with properties that enable evaluation of hierarchical structure extraction without relying on foils must be used. In the present work, we capitalized on the properties of the Fibonacci grammar. This grammar generates self-similar and aperiodic sequences. In these sequences, the learning of one regularity is conditioned by the prior learning of another, lower-level regularity. This makes it possible to evaluate the depth of learning without relying on foils by comparing which regularities the learner has identified. Moreover, due to the self-similar nature of the sequences, the height of the hierarchical structure that can be learned is not limited a priori, which allows for a precise assessment of learning. Secondly, an experimental paradigm that enables continuous tracking of participants' learning performance and avoid explicit grammaticality judgments must be employed. The Serial Reaction Time (SRT) task (Nissen & Bullemer, 1987) precisely meets this requirement. In the SRT task, participants rapidly respond to sequentially presented stimuli, with each response triggering the next stimulus. In the following, we explain in more detail the properties of the Fib grammar strings and explain how their implementation in the SRT task avoids the limitations of the AGL paradigm mentioned in the previous section.

1.2.3.1 The Fibonacci grammar

The Fibonacci grammar (Fib henceforth) belongs to the Lindenmayer formalism (L-systems). In contrast to other classes of grammars, there is no distinction between rewriteable and non-rewriteable symbols in L-systems, and rewrite rules are applied simultaneously to all symbols instead of sequentially from left-to-right in a string (Lindenmayer, 1968; Vitányi & Walker, 1978). These simplifications yield rule systems that generate complex structural patterns. The Fib grammar comprises two rewrite rules:

$$0 \rightarrow 1$$

$$1 \rightarrow 0\ 1$$

The interpretation of this formalism is as follows: every instance of [0] in a sequence is 'rewritten' as [1], and every instance of [1] in the same sequence is rewritten as [01]. By iteratively applying these rules, longer sequences of symbols emerge, each corresponding to a 'generation' of the grammar. This grammar is named after the Fibonacci sequence as the number of points[1] in each

1 Note that in a grammar, the rewriting rules manipulate the "symbols" that belong to an alphabet. The actual realization of symbols can be expressed in various ways without altering their significance. For instance, substituting 0s and 1s with As and Bs has no impact. To refer to the concrete realization of the symbols in the Fib grammar, I will use the term

generation adheres to the sequence (as shown in Fig. 1). Additionally, there is an asymmetry in the distribution of 0s and 1s in each generation, with a greater number of 1s than 0s: the ratio of 1s to 0s approximates the golden ratio (1.618). When examining a sequence generated by the grammar from left to right, two possible transitions exist, one from 0 and the next point, and one from 1 and the next point. The probabilities of these transitions are also asymmetric. The transition from 0 to 1 is deterministic, meaning that 0 is always followed by 1, while the transition from 1 to the next point is probabilistic, with 1 being followed by 0 in 61.8% of cases and by 1 in 38.2% of cases.

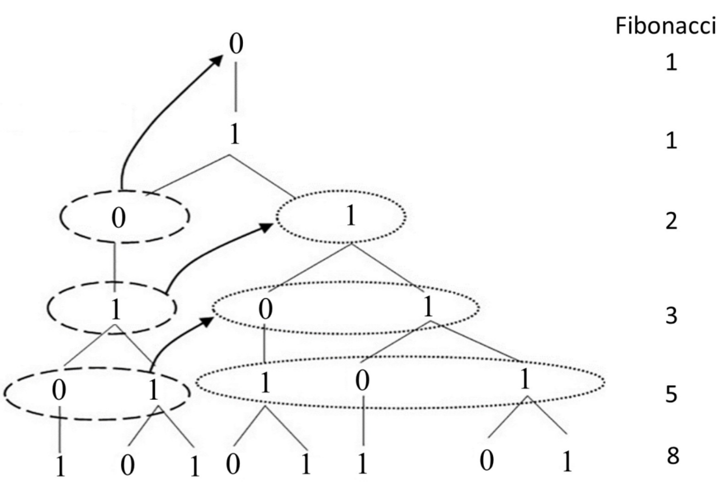


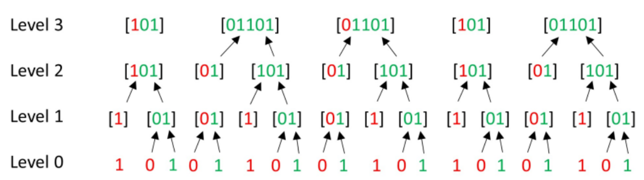
Fig 1. Derivation of the Fibonacci grammar for the first 5 generations. The right column shows the number of symbols at each generation, which maps the Fibonacci sequence. Arrows and circles highlight the hierarchical constituency of the grammar.

The Fib grammar is a particularly suitable candidate for exploring hierarchical structure extraction because it exhibits the following properties: (1) the generated strings are self-similar, (2) the generated strings are aperiodic, (3) the grammar is deterministic, and (4) while presenting complex structural patterns, the generated strings are maximally simplified (i.e., the strings are binary).

The most important aspect of the strings generated by Fib grammar for the present investigation lies in their self-similarity. Because of the recursive aspect of the rewriting rules, each generation in the Fibonacci grammar can be analyzed by breaking it down into two consecutive smaller generations that are inherent constituents of the grammar. For instance, the fourth generation [01101] can be partitioned into the second and third generations [[01][101]], which can be further subdivided into

"point" throughout this thesis.

the first and second generations [[01][1][01]], and then again into the zeroth and first generations [[[0][1]][[1][[0][1]]]]. Therefore, any generation can be visualized as a complex embedding of constituents that exhibit the hierarchical organization of the grammar. Moreover, Fib-generated sequences are scale-free, meaning that the transitional probabilities between points at the surface level are equivalent to those between constituents (see Fig. 2 right panel). Importantly, the points/constituents that surround a deterministic transition at level n always form a larger constituent at level n+1. For instance, at the surface level, 0 is invariably followed by 1, and their concatenation results in the higher-order constituent [01], which is a natural constituent of the grammar. At level 1, the constituent [1] is always followed by the constituent [01], and their combination leads to the higher-order constituent [101]. Therefore, because of the invariance of the transitional probabilities due to the grammar's self-similarity, the distributional properties of Fib provide a scaffold for the parser to access the constituent structure of the grammar. The processing mechanism may begin by merging the points that are connected by a deterministic transition. Subsequently, the higher-order constituents obtained from this merging process can be used to identify deterministic transitions at the next hierarchical level. If this recursive combination process is repeated, it would result in a gradual transformation of the sequence representation into a complex hierarchical structure of embedded constituents (see Fig. 2 left panel).



Non-ambiguous *deterministic transition*	**Ambiguous** *low transitional probability*	**Ambiguous** *high transitional probability*
p([01101]|[101]) = 1	p([01101]|[01101]) = .38	p([101]|[01101]) = .62
p([101]|[01]) = 1	p([101]|[101]) = .38	p([01]|[101]) = .62
p([01]|[1]) = 1	p([01]|[01]) = .38	p([1]|[01]) = .62
p(1|0) = 1	p(1|1) = .38	p(0|1) = .62

Fig 2. Left panel: depiction of the first three hierarchical levels of generation 7 of the Fibonacci grammar. Non-disambiguated points at each level are highlighted in red and disambiguated points in green. To form a new hierarchical level, points that span across a deterministic transition are combined together (this is illustrated by the arrows). The result is a new representation of the string that consists in the combination of points corresponding to natural higher-order constituents of the grammar (illustrated by the brackets). At each level, constituents spanning a deterministic transition can be combined to form an embedded hierarchy. Right panel: transition probabilities between constituents at each level.

The strings generated by the Fib grammar allow us to measure learning online thanks to the following property: points that follow a probabilistic transition at level n can be appear within a constituent that follows a deterministic transition at level n+1. For example, all occurrences of 0s exhibit a probabilistic transition at the surface level, with p(0|1) = .62 and 1-p(0|1) = .38. However, 0s always manifest at level 1 within the constituent [01], and some instances of this constituent follow a higher-order deterministic transition: the constituent [1] is consistently followed by the constituent [01] (p([01]|[1])=1). Therefore, even though all 0s are ambiguous at the surface level, a subset of them are disambiguated at level 1 (i.e., the 0s that follow a higher-order deterministic transition). As a result, the detection of higher-order deterministic transitions serves to disambiguate some of the points that were previously ambiguous at the lower level. Notably, the higher the hierarchical structure, the more points will be disambiguated. Therefore, each hierarchical level corresponds to a specific learning pattern of points: those that remain ambiguous at this level (i.e., non-disambiguated points) and those that are disambiguated at this level.

However, the self-similar aspect of the strings generated by Fib would be of little experimental interest if the distribution of points in the sequences was not also aperiodic. Aperiodicity means that the sequence never exactly repeats itself i.e., there is no linear function that can be used to predict all future points. As a result, there is no hierarchical level that can disambiguate all ambiguous points, a subset of non-disambiguated points which can lead to a new embedding always persists. Due to aperiodicity, it is thus impossible to accurately predict future points using low-level strategies such as detecting recurring patterns.

Moreover, in contrast to a probabilistic sequence, the strings generated by the Fib grammar are entirely deterministic. This makes it possible to generate precise predictions about which anticipation pattern corresponds to a given hierarchical level. Each hierarchical level corresponds to a specific pattern of learning of points: points that are still ambiguous at that level (i.e., non-disambiguated points) and points that are disambiguated at that level and below. This makes it possible to assess the height of the hierarchical structure built by the participants without having to compare performance between grammatical and ungrammatical stimuli, as the learning is evaluated within the sequence itself.

Another advantage of the Fib-generated string is that since sequences are binary, the input presented to participants is maximally simplified while allowing complex structural processing. In contrast, the use of more complex sequences may involve factors that are difficult to control such as prior knowledge, semantic content, or the need to teach participants a lexicon prior to testing (Levelt, 2020; Planton et al., 2021). Finally, the hierarchical structure in the strings generated by Fib is indexed by a single factor, which is the sequential order of presentation of the points. In comparison, other experimental setups typically rely on perceptual cues to index the hierarchical structure of the signal (Franck et al., 2016; Grama et al., 2016; Onnis et al., 2005; Peña et al., 2002; Reeder et al., 2013; van den Bos et al., 2012). The use of perceptual cues can inadvertently introduce associations or biases that may complicate data interpretation. By using sequences that rely only on sequential order to indicate hierarchical structure, this potential complication is avoided, resulting in easier and more reliable interpretation of experimental results.

1.2.3.2 The Serial Reaction Time task

In the present work, we evaluated the learning of Fib-generated string with the Serial Reaction Time (SRT) paradigm (Nissen & Bullemer, 1987). This paradigm allows for the online monitoring of participants' learning performance. In a typical SRT experiment, participants are asked to press a key on a keyboard in response to the appearance of a visual stimulus presented at a specific location on a screen as quickly as possible. Responding to the stimulus causes it to disappear and the next one is presented after a fixed interval. Learning is assessed online: as participants learn the target sequence, they develop expectations about its structure, resulting in improved performance over time (shorter RTs, higher accuracy). By using the SRT task in conjunction with Fib sequences, we can measure learning directly (i.e., without relying on foils) and continuously (i.e., as learning is measured at each trial).

1.3 Research questions

The properties of the Fib grammar coupled with the SRT task provide a novel approach to study hierarchical extraction from sequences while avoiding the pitfalls of the AGL paradigm. The thesis addresses three main questions in regard to the processing of the Fibonacci grammar: (1) Does the cognitive system elaborate a hierarchical representation from Fib-generated strings ? (2) Does one of the formal properties of the Fib grammar, namely the isomorphism between the derivational and the sequential order, play a role in hierarchical structure extraction ? (3) How does the rate at which stimuli are presented affect structure extraction ? To address these questions, a total of 6 experiments have been conducted, which we report in Chapters 2, 3 and 4. In what follows, we briefly outline how these questions are addressed and anticipate the main results reported.

Chapter 2 addresses question (1), which is the precondition for the others, namely whether participants actually elaborate a hierarchical structure when processing this grammar. We hypothesized that the surface properties (i.e., the transitional probabilities) of the sequence would drive the parser in building a hierarchical structure. The mechanism we propose is relatively simple: participants would recursively merge points that span across a deterministic transition (i.e., transitional probabilities equal to 1), and use the output of this process to merge new deterministic transitions between these groups of merged points. This process would result in a representation of embedded constituents. To test this hypothesis, we exposed participants to strings generated by the Fibonacci grammar in a SRT task. Hierarchical elaboration should result in a progressive ability to anticipate locally ambiguous points that are disambiguated at higher hierarchical levels. To test this prediction, we compared the evolution of RTs and accuracy through exposure of points disambiguated at level n to those of non-disambiguated points at the same level. The results indicated that disambiguated points were progressively anticipated better than their non-disambiguated counterparts at levels 0, 1, 2 and 3, suggesting that participants built the structure up to the third hierarchical level. We conducted a second analysis aimed at specifying further whether participants had processed the Fibonacci grammar as a nested structure. If participants rely on deterministic transitions between constituents to anticipate disambiguated points, then the processing of such points would be influenced not only by the level at which they are disambiguated but also by the constituent in which they appear higher in the hierarchy. To this end, we compared different instances of disambiguated points of the same level but occuring at level n+1 either in an ambiguous constituent or in a non-ambiguous constituent. Our results showed that points that had exactly the same transitional probability (which was equal to 1) were processed differently in virtue of their position in the higher-order constituent at levels 1 and 3. This suggest that participants anticipated, at least in part, points on the basis of hierarchical assumptions. In the last part of Chapter 2, we explored in more detail the form of the representations developed by the participants. In particular, we wanted to know whether participants processed constituents as single units without internal structure or as composed of several embedded sub-constituents. The results showed that lower-level constituents were still represented within higher-level constituents, suggesting that participants were not treating constituents as single homogenous units. To conclude, observations from Chapter 2 clearly suggest that participants extracted a hierarchical nested structure from the string generated by the Fib grammar.

Chapter 3 investigates whether hierarchical elaboration is driven by the formal properties of the Fib grammar. More precisely, our specific focus is on investigating the k-points hypothesis (Vender et al., 2020; Krivochen, 2018) according to which participants would be sensitive to the isomorphism between the surface and the derivational properties of the Fib grammar in order to extract its hierarchical structure. According to this hypothesis, the identification of specific points in the grammar, called k-points, would enable the parser to build the local hierarchical structure of the grammar because of their specific structural status. To investigate the cognitive system's sensitivity to the structural status of k-points, Vender et al. (2020) conducted a SRT experiment where they first exposed participants to the Fib grammar and then to an alternative grammar called Skip. The Skip and Fib grammars share some properties: the strings generated are aperiodic and self-similar and the statistical distribution of points is similar. However, from a formal point of view, even though the surface expression of k-points is present in Skip, they do not have any special structural status. The results showed a slowdown in RTs for Skip k-points compared to Fib k-points. These authors concluded that the advantage for Fib k-point would come from their specific formal status. However, this interpretation is questionable because the Skip grammar differs from Fib not only from a formal point of view but also in terms of their surface properties. The difference in the statistical distribution of points between the two grammars is sufficient to account for the results reported by Vender et al. (2020) without having to assume that participants were sensitive to the specific formal properties of Fib. Moreover, the k-point hypothesis seems incompatible with several observations reported in Chapter 2. We therefore wanted to know whether the hypothesis proposed in Chapter 2, namely that participants build a hierarchical structure via the recursive merge of deterministic transitions (referred to as the recursive merge hypothesis), could explain the results of Vender et al. (2020). Insofar as the argument for the k-point hypothesis relies on the comparisons between the Fib and Skip grammar, and insofar as we did not test the Skip grammar in Chapter 2, the purpose of Chapter 3 is to compare the predictions of each hypothesis in the processing of the Fib and the Skip grammar. To this end, we carried out two SRT experiments in which participants were exposed to either the Fib grammar followed by the Skip grammar ("Experiment 1") or the Skip grammar followed by the Fib grammar ("Experiment 2"). This allowed us to test a large number of predictions made by each hypothesis (four for the k-points hypothesis and two for the recursive merge hypothesis). The results showed that in both the Fib and the Skip grammars, participants elaborated a hierarchical structure from the input. This suggests the involvement of at least partially similar mechanisms during the processing of each grammar. The results were also mostly in contradiction with the prediction of the k-point hypothesis whereas the recursive merge hypothesis had the majority of its predictions confirmed. Overall, the results suggest that hierarchical building in Fib is not due to the formal properties of k-point.

In Chapter 4, we explore the impact of the presentation rate on hierarchical structure extraction by manipulating one of the parameters of the SRT task: the Response-to-Stimulus Interval (RSI).

Although this issue has never been directly addressed in the literature, three hypotheses have been proposed to account for the effect of RSI duration in sequence learning in general. According to the Decay hypothesis (Frensch & Miner, 1994; Soetens et al., 2004), RSI duration would affect learning through the decay of stimulus representations in working memory: the shorter the RSI, the better the sequence would be learned. According to the Preparation hypothesis (Willingham et al., 1997), the duration of the RSI would not affect sequence learning as such but rather performance in the SRT task. Learning would be identical across different RSI durations but could only be detected when the RSI is short enough. At long RSI durations, participants would have more time to prepare for the next trial, which would mask the learning effect. Finally, according to the Awareness hypothesis (Cleeremans & Sarrazin, 2007; Destrebecqz & Cleeremans, 2001, 2003; Frensch & Miner, 1994; Huang et al., 2017; Kuhn & Dienes, 2006; Norman et al., 2007; Savalia et al., 2016; Soetens et al., 2004; W. B. Verwey & Wright, 2014; W. Verwey & Dronkers, 2019; Willingham et al., 1997), RSI duration would modulate the implicit/explicit nature of learning: the shorter the RSI, the more learning would be implicit. Under the assumption that abstract knowledge cannot be acquired implicitly, the duration of the RSI would have the effect of restricting the type of representation that can be learned in the SRT task. The knowledge developed would become more and more abstract as the RSI lengthens. To disentangle between these hypotheses, we manipulated the duration of the RSI while using Fib grammar-generated strings as target sequences in three experiments. In Experiment 1, the RSI lasted 1000 ms, in Experiment 2 it lasted 250ms and in Experiment 3 100 ms. According to the Decay hypothesis, the height of the hierarchical structure elaborated by the participants should increase with the shortening of the RSI. The Awareness hypothesis makes the opposite prediction: the height of the hierarchical structure elaborated by the participants should decrease with the shortening of the RSI. Finally, the preparation hypothesis predicts that the duration of the RSI should have no effect on the height of the hierarchical structure elaborated by the participants. The results showed that the height of the hierarchical structure was maximal when the RSI lasted 250 ms and decreased at 1000 ms and 100 ms. These results suggest the existence of an optimal time window for learning. None of the three hypotheses can account for this U-shape, non-linear effect of the RSI. We interpreted these findings through the lens of recent applications of Shannon's information theory to sequence processing (Radulescu et al., 2019, 2021; Shannon, 1948).

1.4 Organisation of the thesis

Chapters 2, 3, and 4 were written as distinct research articles, which resulted in partial overlapping in their introductions. Chapter 2 is published under the title "Finding Hierarchical Structure in Binary Sequences: Evidence from Lindenmayer Grammar Learning" in the journal *Cognitive Science* in January 2023 (Schmid et al., 2023a). Chapter 3 is currently under revision under the title "Uncovering hierarchical structure through statistical dependency" in the journal *PLOS ONE* (Schmid et al., 2023b). Chapter 4 is intended to be submitted in the near future as a research article to an academic journal. Chapter 5 discuss the main results of the experiments further in regard to the main questions addressed in this work. We will highlight limitations of the empirical work and propose directions for future research. To minimize redundancy between chapters, all bibliographical references are gathered at the end of this work.

Chapter 2. Finding hierarchical structure in binary sequences : evidence from Lindenmayer grammar learning

2.1 Introduction

How do humans extract hierarchical structure from a sequentially presented input? This question lies at the core of multiple domains of cognitive psychology and neuroscience. The most prominent is probably language processing where most linguistic theories assume that the sequences that humans produce and remember cannot be reduced to mere associations of consecutive items but must be mentally represented as recursively nested structures (Chomsky, 1957; Lashley, 1951; Simon, 1962). Nested tree structure is a form of representation generated by symbolic rules allowing recursion when they are embedded such that the same element can appear at multiple levels.

There's a plethora of evidence that nested structures are represented and used by adults in sentence processing (e.g., Lewis & Phillips, 2015) as well as in other cognitive domains like mathematical expressions (Maruyama et al., 2012; Monti et al., 2012; Nakai & Sakai, 2014), motor action (Hunt & Aslin, 2001; Martins, Bianco, et al., 2019), musical melody (Koelsch, 2005) and rhythm (Fitch & Martins, 2014; Kotz et al., 2018). Nevertheless, the experimental demonstration of the learning of nested structures in sequence processing has proven difficult and the field of artificial grammar learning (AGL) has produced very few empirical studies showing conclusive evidence (Fitch, 2014; Honing & Zuidema, 2014; Kovács & Endress, 2014; Levelt, 2020).

This difficulty comes from the fact that in the test cases classically used, a sequence can be processed without necessarily building a nested structure as other, possibly simpler ways of representing it can give rise to similar learning performance. Dehaene et al. (2015) proposed a taxonomy of the different types of internal representations that can be generated from a sequence. In particular, they distinguish between two kinds of hierarchical representations: nested representations and algebraic patterns. Algebraic patterns refer to a type of representation where the input is coded as sequential abstract relationships or categories, thus allowing generalization to new exemplars irrespective of their specific identity. For example, the pseudowords "*duduba*" and "*pipiro*" share the algebraic pattern AAB that can be coded as a repetition followed by an alternation. Marcus et al. (1999) showed that at seven months, children were able to generalize this pattern to new unseen pseudowords, suggesting that they had a representation of the AAB rule. Algebraic patterns are hierarchical in the sense that they consist in variables that can take different values. Nevertheless, patterns, even though abstract, are insufficient to account for complex structural dependencies that characterize natural languages, like the subject-verb agreement dependency. For example, in the sentence “[The cats [the car avoided] ran away]” the plural subject (cats) agrees with the verb (ran) irrespective of the intervention of the relative clause (the car avoided). Long-distance dependencies in natural language are impossible to express with a system that only captures local order relations because arbitrarily large materials can intervene between the subject and the verb. In other words, the nesting of constituents where (cats) and (ran) are directly linked is necessary to account for long-distance dependencies.

Many AGL attempts to study the learning of nested structures have focused on the ability to learn and generalize center-embedding (Bahlmann & Friederici, 2006; de Vries et al., 2012; Lai & Poletiek, 2011, 2013; J. L. Mueller et al., 2010). Center-embedding is the nesting of an arbitrary number of phrases into higher-order phrases (e.g., [The cat [the dog chased] ran away]). Context-free grammars (CFG) represent the minimal level in the Chomsky hierarchy because of their unbounded memory that allows the binding of an unlimited number of constituents (Chomsky & Lightfoot, 2002). A well-studied instance of CFG is the a(n)b(n) grammar that generates strings like AB, A[AB]B, A[A[AB]B]B, etc. In order to assess if recursion can be induced by participants after exposure to sequences generated from this grammar, the test contrast is provided by strings generated from a finite state grammar (FSG) like (ab)n. FSGs cannot generate center-embedding because they have no memory; transitions are determined by the current state and the input only.

They are therefore unable to describe nested structures. Fitch and Hauser (2004) compared in a habituation/discrimination task the ability of humans and cotton-top tamarins to discriminate between the a(n)b(n) grammar and the (ab)n grammar. The authors discovered that humans were able to notice the change from one grammar to the other while cotton-top tamarins were not able to discriminate (ab)n from a(n)b(n) after training on a(n)b(n). The results were interpreted as evidence that humans possess a unique ability to induce the hierarchical structure needed to process CFG, while cotton-top tamarins are limited to the processing of less complex grammars.

However, the conclusion that participants can represent a(n)b(n) as a nested structure has been challenged. Perruchet and Rey (2005) noted that it was not necessary to pair As and Bs to discriminate between the two kinds of test strings; a simpler strategy based on counting and detection of repetition could also explain performance. They showed that participants were unable to pair As and Bs in structures involving mirror recursion (center-embedding with systematic pairing of As and Bs that generate strings such as A3[A2[A1B1]B2]B3). Although later studies reported successful learning of mirror recursion under specific conditions (Bahlmann & Friederici, 2006; de Vries et al., 2008, 2012), the authors of these studies all acknowledged that the processing of surface distinctions could also account for performance. This comes from the fact that the ungrammatical test strings necessarily differ in their surface expression from the grammatical string: the correct rejection of an ungrammatical string can therefore also be due to the representation of those surface properties.

Recent work has used fractal stimuli to explore hierarchical processing in the visual modality (Martins et al., 2014, 2015; Martins, Krause, et al., 2019), the auditory modality (Martins et al., 2017, 2020), and in the motor domain (Martins, Bianco, et al., 2019). In this series of studies, participants were performing a completion task on periodic fractals. For example, in Martins et al. (2017) participants were first exposed to three auditory stimuli that were generated by the application of a recursive rule. Participants were then asked to choose between two stimuli the one that followed the rule at the higher hierarchical level. Each application of the rule added a hierarchical level to the existing stimulus. Each hierarchical level consisted of three notes that formed an ascending contour. The application of the recursive rule superimposed on each note of the preceding level three shorter higher pitch notes that also formed an ascending contour. For example, the first stimulus was a low-pitch note with a duration of 7.3 sec (level 1). The second stimulus (level 1 + level 2) superimposed three shorter medium pitch notes on the low pitch note of level 1. The third stimulus (level 1 + level 2 + level 3) superimposed nine shorter high pitch notes on each of the medium pitch notes of level 2. The authors found that participants were able to select the correct continuation when presented along with different foils, and interpreted this result as an indication that participants were able to apply rules to new hierarchical levels. However, these results do not demonstrate that rules were embedded because it was sufficient to apply the rule only to the highest hierarchical level to solve the task. Indeed, a rule of the type "the notes follow an ascending pattern" was enough to reject the foils because level 3 of each foil violated this rule. In other words, it was not necessary to apply the rule simultaneously at all the hierarchical levels to succeed.

As we have seen, it has proven challenging to create foils that allows to distinguish between learning based on surface regularities from learning based on higher-order structural properties in the habituation/discrimination paradigm. Furthermore, the presentation of ungrammatical strings may contaminate participants' mental representations throughout the testing phase. To avoid these difficulties, one should be able to assess learning without having to present ungrammatical strings to participants. To this end, the grammar should generate sequences in which the learning of one regularity is conditioned by the learning of another, lower-level regularity. This makes it possible to evaluate the depth of learning by comparing which regularities the learner has identified. Assessing learning of such a grammar that contains its own test can be done with a procedure that measures the evolution of performance throughout the task, avoiding the use of ungrammatical strings and explicit grammaticality judgments. The serial reaction time (SRT) paradigm (Nissen & Bullemer, 1987) allows such on-line monitoring of the participants' learning performance. In the SRT task, participants respond as quickly as possible to successively presented stimuli, usually by pressing response keys. Each response triggers the presentation of the next stimulus, to which participants respond anew. Learning typically manifests by a reduction in reaction times and is expected to take place when a given trial is subject to anticipation.

Only a few studies have made use of this paradigm to explore the learning of hierarchical structure and for most of them, the kind of knowledge developed by participants involves algebraic patterns and not nested structures. Koch and Hoffman (2000) were the first to report evidence suggesting sensitivity to higher-order properties of sequences in SRT. Participants were presented with sequences consisting of 6 different digits. The sequences were periodic and 24 digits in length. The participants' task was to respond to the digit presented on the screen with one of the six response keys. The authors manipulated the relational structure of the sequences. In the third experiment, the highly structured sequences were composed of four pairs of three elements that followed two relational patterns. The first two pairs corresponded to a mirror relationship of an ascending and descending order (e.g., 123-321) and the last two pairs corresponded to a transposition (e.g., 123-234). Participants in this condition therefore saw a sequence like 123-321-456-654-123-234-345-456 (e.g., mirror, mirror, transposition, transposition). The unstructured sequences were created by the permutation of the triplets in such a way to break the relational patterns while keeping the statistical distribution identical (e.g., 123-345-456-123-234-321-456-654). The results showed a greater decrease in reaction times for participants in the structured than in the unstructured condition, suggesting that they were sensitive to the sequences' higher-order relational structure. The participants thus went beyond the surface statistical properties and seem to have organized the sequence according to relational patterns. However, an algebraic rule like "two mirror relations followed by two transposition relations" is actually sufficient to account for the results: it is thus not necessary to assume that the representation developed by the participants corresponds to a nested structure in which an algebraic rule is nested within another algebraic rule, since the relational patterns were not embedded in multiple levels.

In a slightly different task, the discrete sequence production task (DSP), Verway and Wright (2014) trained participants by repeatedly presenting them with short sequences of six elements, each was associated with the location of an illuminated square. During training, each sequence was presented with one of the six elements positioned in a random location, while all other elements occupied a position following a pattern, which could not be extracted from a single sequence, but required combining positional information across sequences. In the test phase, participants were presented with the sequence without deviations (i.e., the ‘true’ but never seen sequence) as well as an unfamiliar sequence (i.e., a sequence where the order of elements never matched the training phase). Participants were faster in the no-deviation sequence than in the unfamiliar sequence, although they did not practice either during the training phase. This suggests that during the training phase, participants extracted probabilities related to the order of appearance (i.e., the probability that an element appears in position 1, position 2, etc.) and combined that information into a representation capturing the underlying pattern of the sequence. Although those results demonstrate learning of an algebraic pattern, like in the study of Koch and Hoffman (2000), they do not attest to learning of nested structures.

To our knowledge, only one SRT study reported results suggesting the use of nested structures, which is that from Hunt and Aslin (2001). These authors presented probabilistic sequences in a visual SRT task. The sequences were presented by illuminating buttons occupying different spatial positions. In their Experiment 3, the sequence consisted of 4 pairs of elements where the transitional probability from the first to the second element was 1, so the second element of a pair could always be anticipated with certainty by the participants. On the other hand, the transition between pairs was governed by the following probabilities: pairs A and B were each followed in 50% of the cases by pair C and in the remaining 50% by pair D. Pairs C and D were each followed in 25% of the cases by pair A and in 25% of the cases by pair B. Pair C was followed in 50% of the cases by pair D and pair D in 50% of the cases by pair C. An additional restriction was that when pairs C and D were contingent, the next pair had to be either A or B (thus prohibiting alternating CDC or DCD). The authors observed that some participants became sensitive to the cumulative probability of the two most frequent pairs. When pairs C and D were contingent, reaction times for the second element of the pair in position 2 were faster than those for the second item of the same pair when it was in position 1. Since the transitional probability was always 1 for the second element of a pair, the effect can be explained only if participants have acquired the knowledge that the transition between elements of a pair is embedded in the transition between pairs. This embedding of transition seems more in line with a nested representation than a representation of an algebraic pattern; however, this interpretation has some limitations. First, only 3 participants out of 10 showed the effect. Second, the alternation CDC and DCD being prohibited, the transition following CD or DC was at chance level (50% A and 50% B). Thus, the design of the materials prevented determining if participants nested more than one relation, that is, if the transitions between pairs were themselves embedded into transitions between multiples pairs. Nevertheless, the results suggest that transitional information is sufficient to bootstrap the construction of nested representations.

In a recent study, Planton et al. (2021) went further and explored if a simple form of temporal sequence could give rise to nested representations. One of the simplest forms of temporal sequences are binary sequences, and unlike more complex sequences like music or natural language, they have the advantage of allowing maximal control of the input presented to the participants. This apparent simplicity however preserves the possibility of creating highly complex sequences, which can be expressed as nested tree structures. The authors presented short binary sequences in a violation detection task. After an exposure phase, altered sequences that deviated by one item from the initial sequences were presented to the participants. The participants' task was to report as quickly as possible if they detected a violation. In order to vary the complexity of the sequences, the authors developed a formal language containing a limited number of primitive instructions that could generate any binary sequence. This allowed them to characterize each binary sequence in terms of *Kolmogorov Complexity*. Kolmogorov complexity is a theoretical measure where the complexity of a sequence is equal to the size of the shortest computer program that can generate it. Thus, the complexity of a sequence was defined by the minimal number of primitive instructions needed to generate it in the proposed language. The more the complexity of a sequence increases, the more its most compressed representation requires the use of instruction nesting. The authors therefore wanted to know if the participants' sequence representations were compressed in a similar way. To separate the part of the performance explained by this compression process and the part that can be attributed to the learning of transitional probabilities, the authors also measured in each test sequence the *Shannon surprise* induced by the deviant stimuli. Shannon surprise (Shannon, 1948) measures the degree of uncertainty of observing an item given the history of previous items and thus reflects statistical learning. Since surprise is independent of complexity (it varies with the position of the deviant within a sequence and is insensitive to sequence complexity that characterizes a sequence as a whole), if participants process only the transitional probabilities of the sequences, the degree of surprise of the deviant stimuli should be the only predictor of performance. Conversely, the use of compression by participants should result in a significant portion of the variance being explained by the degree of complexity of the sequences. The results showed that both surprise and complexity were significant predictors of performance suggesting that compression occurred along with statistical learning. This finding demonstrates that statistical learning is insufficient to fully account for sequence processing: even when processing sequences as simple as binary sequences, participants recode the sequence using a recursive compression algorithm. However, this study did not assess the degree of compression of the participants. Indeed, sensitivity to complexity, demonstrated by slower violation detection times in the most complex sequences, does not imply that participants have compressed the sequence to the maximum, nor that the primitive instructions of their formal language correspond to the mental operations of the participants. Our study aims to go further by trying to characterize more precisely the mechanism used by the participants to compress the signal.

2.1.1 Present study

The purpose of the present study is to evaluate, with the SRT paradigm, if participants represent binary sequences of events as nested structures. In theory, recursive compression algorithms allow an infinite number of hierarchical levels. This is obviously not the case for human whose processing capacity is finite, limiting the number of hierarchical levels it can represent. Nevertheless, this limit cannot be defined a priori and can vary from one participant to another. Thus, predefining in advance the hierarchical structure of a sequence and setting a maximum number of levels does not allow for finely evaluating the hierarchical depth reached by the participants. We avoided this problem by using sequences generated by the Fibonacci grammar that are self-similar and aperiodic. The investigation of hierarchical processing with sequences having these two properties has several advantages. First, the self-similar character of the sequences does not limit a priori the hierarchical depth, which is theoretically infinite[1]. Second, the aperiodic character of the sequences means that no matter how deep the hierarchical representations are, they will necessarily be incomplete and will only explain part of the signal. Thus, the part not explained by the hierarchical structure corresponds to the maximum hierarchical level reached. In this way, it is not necessary to compare performance between grammatical and ungrammatical stimuli because the learning is evaluated within the sequence. Crucially, the linear distribution of units (henceforth referred to as *points*) in the sequences is aperiodic, meaning that there is no linear function that can be used to linearly predict *when* a point will occur. This prevents the use of low-level strategies like detecting recurring patterns.

The sequences we will use are generated by a grammar derived from the Lindenmayer formalism (L-systems). These grammars show interesting properties: there is no distinction between rewriteable and non-rewriteable symbols, and rewrite rules apply simultaneously to all symbols[2]

1 Note that the hierarchical depth can of course only be infinite for an infinite chain. In the present study, the presented sequences were 233 points long, and had potentially up to 12 hierarchical levels, which is presumably well beyond the processing capacity of the cognitive system.

2 Formally, the rewriting rules of a grammar operate on the "symbols" of an alphabet. The expression of the symbols (i.e. their actual realization) can however vary. For example, 0s and 1s can be replaced arbitrarily by As and Bs without any impact. In this article, we use the term "point" to refer to the actual realization of the symbols of the Fibonacci grammar.

rather than sequentially from left-to-right in a string (Lindenmayer, 1968; Vitányi & Walker, 1978). Because L-systems do not distinguish rewriteable from non-rewriteable symbols, rule systems are simplified, but still produce complex structural patterns. One instantiation of L-systems used in AGL paradigms is the so-called Fibonacci grammar, which consists in two rewrite rules (Geambaşu et al., 2016; Saddy, 2009; Shirley, 2014):

$$
\begin{gathered}
0 \rightarrow 1 \\
1 \rightarrow 0\ 1
\end{gathered}
$$

The interpretation of such a formalism is very simple: every instance of [0] in a sequence must be 'rewritten as' [1], and every instance of [1] in the same sequence must be rewritten as [01]. Applying these rules over and over again generates longer and longer sequences of points, each of which corresponds to a 'generation' of the grammar. The name of this grammar comes from the fact that the number of points in each generation actually follows the Fibonacci sequence (Fig. 1C). Moreover, in each generation, the distribution of 0s and 1s is asymmetric, with more 1s than 0s: the ratio between the number of 1s and 0s approximates the golden ratio (1.618). If we consider a sequence (i.e., a string generated by the grammar) from left to right, two transitions are possible (from 0 and the next point and from 1 and the next point) and the probability of those transitions is also asymmetric. The transition from 0 to 1 is deterministic: 0 is always followed by 1. The transition from 1 to the next point is probabilistic: 1 is followed by 0 in 61.8% of the cases and by 1 in 38.2% of the cases.

The most important property of this grammar with respect to our research question is its self-similarity. Each generation of this grammar constitutes by definition a natural constituent (Krivochen et al., 2018). Because of the recursive nature of the generative process, any generation is the concatenation of the two previous generations (Fig. 1C). This means that any generation can be parsed with two consecutive smaller generations that are natural constituents of the grammar. For example, generation 4 [01101] can be divided into generations 2 and 3 [[01][101]], which can be further divided into generations 1 and 2 [[01][1][01]], which can (trivially) be further divided into generations 0 and 1 [[[0][1]][[1][[0][1]]]]. Thus, any generation can be seen as a multiple embedding of constituents reflecting the hierarchical structure of the grammar. Transitions in the Fibonacci grammar are scale-free: the transitional probabilities between points at the surface level are identical to the transitional probabilities between constituents (Fig. 1A right panel). Crucially, points/constituents surrounding a deterministic transition at level n always form a bigger constituent at level n+1. For example, at the surface level, 0 is always followed by 1 and the concatenation of these two points results in the higher-order constituent [01], which is a natural constituent of the grammar. At level 1, the constituent [1] is always followed by the constituent [01] and their concatenation results in the higher-order constituent [101]. Thus, because of the grammar's self-similarity, transitional probabilities at each level provide the parser a way to access the constituent structure of the grammar. The processing mechanism may start by merging the points linked by a deterministic transition, and then use the output of this process, i.e., the higher-order constituents, to detect the deterministic transitions at the next hierarchical level. This process of recursive combination would progressively transform the representation of the sequence into a complex hierarchical structure of embedded constituents (Fig. 1A left panel).

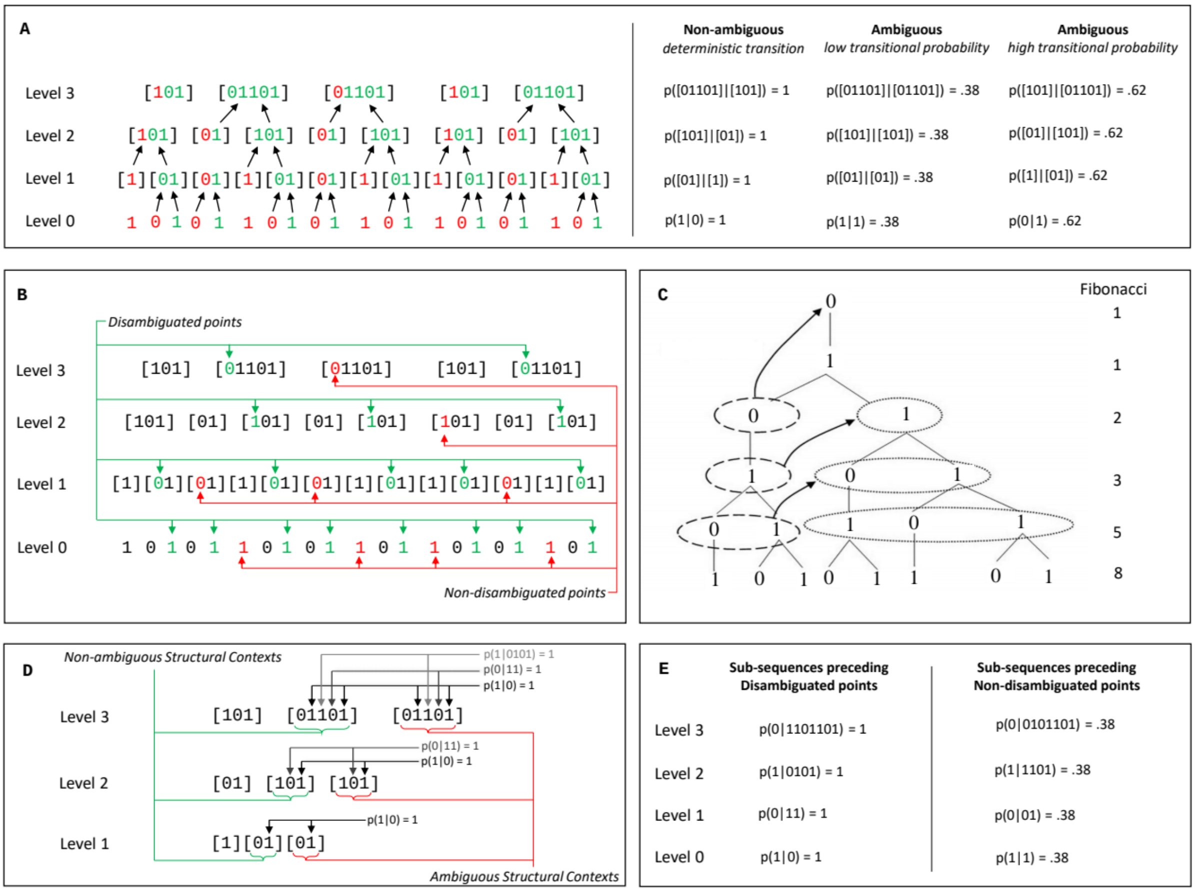


Fig 1. (A) Left panel: depiction of the first three hierarchical levels of generation 7 of the Fibonacci grammar. Non-disambiguated points at each level are highlighted in red and disambiguated points in green. To form a new hierarchical level, points that span across a deterministic transition are combined together (this is illustrated by the arrows). The result is a new representation of the string that consists in the combination of points corresponding to natural higher-order constituents of the grammar (illustrated by the brackets). At each level, constituents spanning a deterministic transition can be combined to form an embedded hierarchy. Right panel: transition probabilities between constituents at each level. (B) Disambiguated points (green) and non-disambiguated points (red) for each hierarchical level for generation 7 of the Fibonacci grammar. In the present study, we used generation 12 of the Fibonacci grammar that consists in 233 points. We did not illustrate this generation due space limitation, but the rationale is identical. (C) Derivation of the Fibonacci grammar for the first 5 generations. The right column shows the number of symbols at each generation, which maps the Fibonacci sequence. Arrows and circles highlight the hierarchical constituency of the grammar. (D) Structural contexts at levels 1, 2 and 3. Green bars point to the constituents in non-ambiguous structural contexts at each level and red bars point to the same constituents when in ambiguous structural contexts. Arrows illustrate the fact that, with the exception of the first point, points that occur inside constituents have the same transitional probability regardless if the constituent is in an ambiguous or non-ambiguous structural context. (E) Transitional probabilities for disambiguated and non-disambiguated points at each level given the sub-sequence that precedes them. We see that the transitional probability of the sub-sequence that precedes a disambiguated point is equal to 1 whereas the transitional probability of the sub-sequence that precedes a non-disambiguated point is equal to .38.

This leads to an interesting observation: points that follow a probabilistic transition at level n can appear inside a constituent that follows a deterministic transition at level n+1. For example, all 0s follow a probabilistic transition at the surface level: $p(0|1) = .62$ and $1\text{-}p(0|1) = .38$. However, 0s always appear at level 1 in the constituent [01] and some instances of this constituent follow a higher-order deterministic transition: the constituent [1] is always followed by the constituent [01] ($p([01]|[1])=1$). Thus, although at the surface level all 0s are ambiguous (i.e. they follow a probabilistic transition) a subset of them are *disambiguated* at level 1 (i.e. the 0s that follow a higher-order deterministic transition). Therefore, the detection of higher-order deterministic transitions serves to disambiguate some of the points that were ambiguous at the lower level. Crucially, the higher the hierarchical structure is, the more ambiguous points will be disambiguated. Nevertheless, due to the aperiodicity of the string, there will always remain a subset of non-disambiguated points that can lead to new embedding, no matter the depth of the hierarchy. Thus, each hierarchical level corresponds to a specific learning pattern of points: points that are still ambiguous at this level (i.e. non-disambiguated points) and points that are disambiguated at this level and lower levels.

Structural processing in the Fibonacci grammar has already been explored via the classical AGL paradigm (Geambaşu et al., 2016, 2020). However, these studies have run into the problem inherent to the habituation/discrimination paradigm of creating non-grammatical test strings that respect the surface properties of grammar. In a first study, Geambaşu et al. (2016) found that participants exposed to the Fibonacci grammar were unable to distinguish between grammatical and ungrammatical strings, and attributed that failure to the fact that some of the foils were in fact Fib-grammatical (i.e., they were possible subsequences of the Fibonacci grammar). In a follow-up study using a different set of non-grammatical test strings, Geambaşu et al. (2020) found that participants were able to discriminate them from grammatical strings, and concluded that the grammar was successfully learned. However, closer inspection shows that 16 of the 18 foils contained the non-grammatical sub-sequence [01010], which is impossible in the Fibonacci grammar. Hence, participants may have rejected the foils on the basis of a low-level strategy without having learned the Fibonacci grammar. Two other studies (Vender et al., 2019, 2020) explored the Fibonacci grammar by way of an SRT task: a sequence of blue and red dots generated by the Fibonacci grammar was presented to the participants whose task was to press the left or right button corresponding to the color of each dot. Sequences of dots were implemented in a Simon task: dots appeared to the left or to the right side of the screen, such that the colored dot sometimes appeared to the opposite side of the corresponding key. Such incongruent trials occurred every sixth trials. The Simon task was introduced to make the task less repetitive for participants. In the 2020 study, the authors added a final block within which the order of appearance of stimuli followed an alternative grammar called Skip, which has similar surface properties to Fib: 0 is always followed by 1 ($p(1|0) = 1$), the sub-sequence 11 is always followed by 0 ($p(0|11) = 1$) and the first order transitional probabilities are relatively similar: $p(0|1) = .73$ and $p(1|1) = .27$ but differ from the latter from a formal point of view. The authors proposed that within the Fibonacci grammar, the identification of certain points, called k-points, would allow the reconstruction of the local hierarchical structure of the sequence due to their specific structural status. Indeed, the distance between two k-points exactly mirrors the transitional probability of the minimal units of the sequence (see Krivochen et al., 2018 for a detailed explanation). Linearly, k-points are the last 1 of the 3-gram [011] and correspond to the constituent [1] of level 1 (shown in Fig. 1-A left panel) whose transitional probability is $p(1|1) = .38$. In Skip, although the surface expression of the k-points is present (Skip has the 3-gram [011]), their identification would not allow the reconstruction of the local hierarchical structure because the distance between them does not mirror the statistical distribution of minimal units. In other words, in contrast to Fib, the self-similarity of Skip does not allow to extend the local statistical regularities at a higher hierarchical level. Vender et al. (2020) found faster processing for the last 1 of the 3-gram [011] in Fib blocks than in the Skip block. They interpreted this as evidence that participants had granted a special status to k-points, suggesting that they partially reconstructed the hierarchical structure of the Fibonacci grammar.

However, a more detailed analysis of the sequences generated by the Skip grammar shows an inversion of the second order transitional probabilities. In Skip, k-points have a second order conditional probability of $p(1|01) = .36$ while in Fib it is equal to $p(1|01) = .62$. Thus, the slower processing observed for the last 1 of the 3-gram [011] in Skip block could also be explained by participants becoming sensitive to the fact that 01 is more frequently followed by 0 than by 1. The effect can therefore also be explained by '*flat*' statistical learning processes. Moreover, the Simon task introduces a factor that occurs periodically (i.e. incongruent trials occurred every sixth trials); Fibonacci grammar being aperiodic, incongruent trials are not distributed evenly in the sequence, which makes the impact of this factor difficult to evaluate.

In the present study, we implemented Fibonacci sequences in an SRT task, thus avoiding the need to create non-grammatical Fib-strings (like in Geambaşu et al., 2016, 2020). In contrast to Vender et al. (2019, 2020), dots were presented in the center of the screen, to avoid the interfering congruency factor introduced by the Simon task. Importantly, we developed new analyses, substantially different from those conducted in these 4 papers, which allowed us to evaluate hierarchical learning within the Fibonacci grammar without having to compare the performance of participants to another grammar or to a random block. Sequence learning in the SRT task is traditionally assessed by inserting a so-called "transfer block" at the end of the experiment in which trials follow a random order or an alternative sequence. A slowdown in the transfer block relative to the block that precedes it is interpreted as indicating that participants have acquired the target sequence (Schwarb & Schumacher, 2012). However, when it comes to interpreting the origin of a slowdown in the transfer block this methodology encounters the same limitation as the habituation/discrimination paradigm. The slowdown can be either due to a change in surface properties or to a change in more abstract properties. The use of the Fibonacci grammar aims precisely at avoiding this problem because it allows us to evaluate the learning during the processing without having to compare the performance to an alternative sequence. Our conceptual framework critically diverges from Vender et al. (2020) in that rather than hypothesizing that the parser extracts some formal properties of the

Fibonacci grammar (k-points), we hypothesize that it proceeds through recursively merging points that span across deterministic transitions, and then using the output of this process to merge new deterministic transitions between groups of points, resulting in the progressive building of a hierarchical structure. Participants may also develop knowledge of formal properties of the Fibonacci grammar, however this question is beyond the scope of the present study.

We carried out two analyses to assess whether participants built a hierarchical structure from the Fibonacci grammar through the recursive combination of points/constituents surrounding deterministic transitions. The first analysis (*Processing of hierarchical structure*) explored whether disambiguated points (i.e. points following a higher-order deterministic transition) were anticipated better than non-disambiguated points (i.e. points following a higher-order probabilistic transition). To this end, we compared reaction times and accuracy for points disambiguated at a particular hierarchical level to points not disambiguated at the same level (Fig. 1B). Hierarchical processing should result in a larger decrease in reaction times and better accuracy for disambiguated points compared to non-disambiguated points. We do not have any prior expectation with respect to how many levels the participants might reach. We will therefore evaluate each level successively until the effects disappear at the group level (see Fig. 1A left panel for levels descriptions). In order to control for frequency effects that could be due to the asymmetry of the sequence (1s being more frequent than 0s), we compared, for each hierarchical level, only 1s to 1s and 0s to 0s. Anticipating the results, we found evidence of learning at levels 1, 2 and 3 but not at level 4 (which is why this level is not presented in Fig. 1A and Fig. 1B).

The second analysis (*Processing of hierarchical constituency*) aimed at specifying further whether participants have processed the Fibonacci grammar as a nested structure. To this end, we explored the influence of the constituent structure at level n on the processing of disambiguated points at level $n-1$. This analysis is a logical continuation of the first: if participants use deterministic transitions between constituents to anticipate disambiguated points, then the processing of a disambiguated point should depend not only on the level at which it is disambiguated but also on the constituent in which it appears higher in the hierarchy. If we examine closely the constituents of each level, we see that the first position (from left to right) is always occupied by either a disambiguated or a non-disambiguated point (Fig. 1A left panel) whereas the following positions are composed of points disambiguated at the previous levels. Crucially, the remaining positions of the constituent following a deterministic transition and of the constituent following a low probabilistic transition (Fig. 1A right panel) are occupied by points disambiguated at the same levels (Fig. 1D). In other words, a point disambiguated at level n can appear at level n+1 in either a constituent that follows a deterministic transition or in a constituent that follows a probabilistic transition, while the composition of the constituents is identical (except for the point in the first position). Thus, the same disambiguated point appears higher in the hierarchy subsumed in a different structural context. We refer to the condition where a disambiguated point appears at a higher level inside a constituent that follows a deterministic transition as a *non-ambiguous structural context* and to the condition where a disambiguated point appears at a higher level in a constituent following a probabilistic transition as an *ambiguous structural context* (Fig. 1D). If the system is sensitive to the hierarchical constituency of the sequence, disambiguated points appearing at the upper level in a non-ambiguous structural context should be processed faster than the same disambiguated points appearing in an ambiguous structural context. Anticipating the results, we found a significant processing advantage for points occurring in non-ambiguous structural contexts compared to points occurring in ambiguous structural contexts at levels 1 and 3.

2.2 Methods

2.2.1 Participants

One hundred seventy-four students (33 men and 141 women; mean age 22.8 years old) participated in the experiment. They were recruited either from an introductory psycholinguistics course from the university of Geneva or through announcements at the University of Geneva. All participants reported normal or corrected-to-normal vision.

2.2.2 Materials

The training sequence was composed of two elements and had a length of 50. The order was pseudo-randomized and elements had the same frequency. The training sequence included multiple non-grammatical sub-sequences such as 00 or 111. The longest Fib-grammatical sub-sequence had a length of 4. In the experimental blocks, the sequence consisted of generation 12 of the Fibonacci grammar which has 233 points. Each block corresponded to the full generation.

2.2.3 Design and procedure

Each trial consisted of a red or blue circle 100px in diameter presented at the center of the screen which correspond, respectively, to 0 and 1 in a string generated by the Fib grammar. The circles disappeared after the response of the participant, or after 1200ms, if no response was given. The response-to-stimulus interval lasted 500 ms. Participants were instructed to press as quickly as possible the button corresponding to the color of the circle they saw on the screen (X=blue, N=red). Keys X and N were chosen because they had a similar position on QWERTZ and AZERTY keyboards. No information about the grammar was given. The experiment started with a training block that was identical for all the participants. During the training block, when the participants made an error, the experiment stopped and a message appeared to remind them the color – key association, the experiment resumed after 3000ms. In the experimental blocks, no message appeared when they made an error. After the training block, participants did 5 experimental blocks of 233 trials. The experiment was conducted online on the website Testable (https://www.testable.org/) (Rezlescu et al., 2020). Pre-testing showed that the error rate in the task was extremely low, which is not surprising given the simplicity of the task, so the emphasis on speed alone was intended to increase the error rate and avoid ceiling effects. Participants were asked to perform the experiment in a quiet environment where they could not be disturbed. Instructions were displayed on the screen and participants had to click on a button to start the experiment. The experiment lasted approximately 25 minutes.

2.2.4 Data analyses

Four participants were removed due to technical failures. We also removed participants who had an error rate superior to 3 *SD* to the mean error rate in at least one block. This led to the removal of 11 additional participants. Due to an error in the experiment code, the data of the training block was not recorded. Reaction times and accuracy were both modelled as dependent variables. We removed from the analysis all the trials where participants did not respond after 1200 ms (699 trials). For the analysis of reaction times, only trials with a correct answer were included. Homoscedasticity and normality were checked by visual inspection of residual plots. Data from the remaining 159 participants were analyzed with linear mixed-effects models as implemented in the lme4 package for R (Bates et al., 2014; R Core Team, 2022).

For the analysis *Processing of hierarchical structure*, models included two fixed-effect factors and their interaction: *Exposure, Ambiguity,* and *Exposure*Ambiguity*. Exposure was treated as a continuous variable with a value of 0 for trials of the 1st experimental block, and of 1, 2, 3 and 4 for trials of the 2nd, 3^{d}, 4th and 5th blocks. This factor being continuous, it allowed us to have only one estimate which represents the evolution (i.e., the slope) of performance throughout the experiment across all participants. Ambiguity is a discrete variable contrasting disambiguated and non-disambiguated points and operationalized differently depending on the level at which its effect is explored (it is labeled *Ambiguity level*$_n$ according to the level at which it has been operationalized). The modality “Non-disambiguated” of the factor *Ambiguity level*$_n$ was always set as the intercept of the models. As random effects, the models had intercepts for *Participants*. P-values were calculated by way of the Satterthwaites’s approximation to degrees of freedom with the lmerTest package (Kuznetsova et al., 2015). We conducted separate analyses for RTs and accuracy instead of using a composite score because there is no consensus in the literature on the optimal method of calculation (Liesefeld & Janczyk, 2019; Vandierendonck, 2017, 2018). Moreover, composite measures that integrate RTs and accuracy cannot be calculated per trial but only per condition (for each participant). Since the factor *Ambiguity* is nested within blocks (i.e., each block contains several disambiguated and non-disambiguated points of the same level), using a composite score would drastically reduce the number of observations per participant and thus the statistical power of the analyses.

For the analysis *Processing of hierarchical constituency*, models included two fixed-effect factors and their interaction: *Exposure*, *Structural context*, and *Exposure*Structural context*. Structural context is a discrete variable contrasting disambiguated points that appeared at the next level in constituents that either followed a deterministic transition (Non-ambiguous) or a probabilistic transition (Ambiguous). This variable is operationalized differently depending on the level at which its effect is explored (it is labeled *Structural context level*$_n$ according to the level at which it has been operationalized). The same mixed models were ran as in previous analysis, including *Structural context* and *Exposure* as fixed factors and the modality "Ambiguous" of the factor *Structural context level*$_n$ was always set as the intercept. Since at each level, the first point of a constituent is either a disambiguated or a non-disambiguated point we excluded those first points when we computed the mean RTs and accuracy of the constituents (Fig. 1D). At level 1, the constituent of interest is [01] and contains 2 points. This constituent is in a non-ambiguous structural context when it is preceded by [1] but in an ambiguous structural context when preceded by [01]. Since the first point of [01] can be either a disambiguated or a non-disambiguated point, we have included in the level 1 analysis only the second point of constituent [01] (i.e., the 1). We excluded from the RT analyses all the constituents containing at least one error. At level 2, the constituent of interest is [101] and contains 3 points. It is in a non-ambiguous structural context when it is preceded by [01] and in an ambiguous structural context when it is preceded by [101]. We have included in the analysis only the last two points of constituent [101] (i.e., 01) for the same reason explained above. In the RT analysis, we first excluded all constituents containing at least one error. Insofar as the distribution of 0s and 1s is identical in each modality of the factor *Structural context* (i.e. there is exactly one 0 and one 1 in both the non-ambiguous and the ambiguous structural context at level 2), there is no more asymmetry between the number of 0s and 1s. We thus calculated for each occurrence of the constituent [101] the mean of the last two points and took this measure as dependent variable. For analyzing accuracy, we computed the mean number of correct answers for the last two points of [101] (i.e., the two disambiguated points that appeared in both structural contexts) and divided it by 2 in order to have a value that ranged from 0 to 1 (we did not consider the first point of [101] because it could either be disambiguated or non-disambiguated point depending on the structural context). At level 2, the accuracy value for the constituent was either 1 (no error), 0.5 (1 error) or 0 (2 errors). At level 3, the constituent of interest is [01101] and contains 5 points. It is in a non-ambiguous structural context when it is preceded by [101] and in an ambiguous structural context when it is preceded by [01101]. We have included in the analysis only the last four points of constituent [01101]. To analyze RTs, we first excluded all constituents containing at least one error. We then calculated for each occurrence of the constituent [01101] the mean of the last four points (i.e., 1101) and took this measure as dependent variable. For analyzing accuracy, we followed the same logic as in level 2 but with the constituent [01101]. We computed the mean number of correct answers for the four disambiguated points that appeared in both structural context and divided it by 4 in order to have a value that ranged from 0 to 1. At level 3, the accuracy value for the constituent could either be 1 (no error), 0.75 (1 error), 0.5 (2 errors), 0.25 (3 errors), or 0 (4 errors).

We first explored if participants were sensitive to the surface statistical properties of the sequence, corresponding to level 0, and then if they were able to detect the higher-order deterministic transitions at levels 1-4 (see Fig. 1B). We then explored if participants were sensitive to the constituent structure of the grammar by comparing, at each level, disambiguated points occurring in different structural contexts (see Fig. 1D). Finally, we analyzed performance at the individual level to more finely explore the effect of structural context at level 3 found at the group level.

2.3 Results

2.3.1 Processing of hierarchical structure

2.3.1.1 Processing of surface statistical regularities (Level 0)

Analyses of reaction times showed a main effect of *Exposure* (β = -21.53, *SE* = 0.25, *t* = -87.23, *p* < .001) with a mean reduction of reaction times of 86 ms from block 1 to block 5. There was also a main effect of *Ambiguity* $level_0$ (β = -57.45, *SE* = 0.73, *t* = -78.52, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 57 ms. The interaction *Ambiguity* $level_0$* *Exposure* was also significant (β = -14.16, *SE* = 0.50, *t* = -28.48, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block1 - block5}$ = -106 ms) than non-disambiguated points ($M_{block1 - block5}$ = -49 ms) ($M_{block1 - block5}$ indicates the mean difference between blocks 1 and 5). Results are shown in Fig. 2.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.06, *SE* = 0.01, *z* = -6.302, *p*< .001) with a mean reduction of accuracy of 1 % from block 1 to block 5. There was also a main effect of *Ambiguity* $level_0$ (β = 2.26, *SE* = 0.04, *z* = 57.80, *p*< .001) with higher accuracy for disambiguated points (*M* = 0.98) than for non-disambiguated points (*M* = 0.90). The effect of *Exposure* significantly interacted with *Ambiguity* $level_0$ (β = 0.23, *SE* = 0.03, *z* = 8.354, *p*< .001) with accuracy increasing for disambiguated points over exposure ($M_{block1 - block5}$ = 0.006) and decreasing for non-disambiguated points ($M_{block1 - block5}$ = -0.037). Results are shown in Table 1.

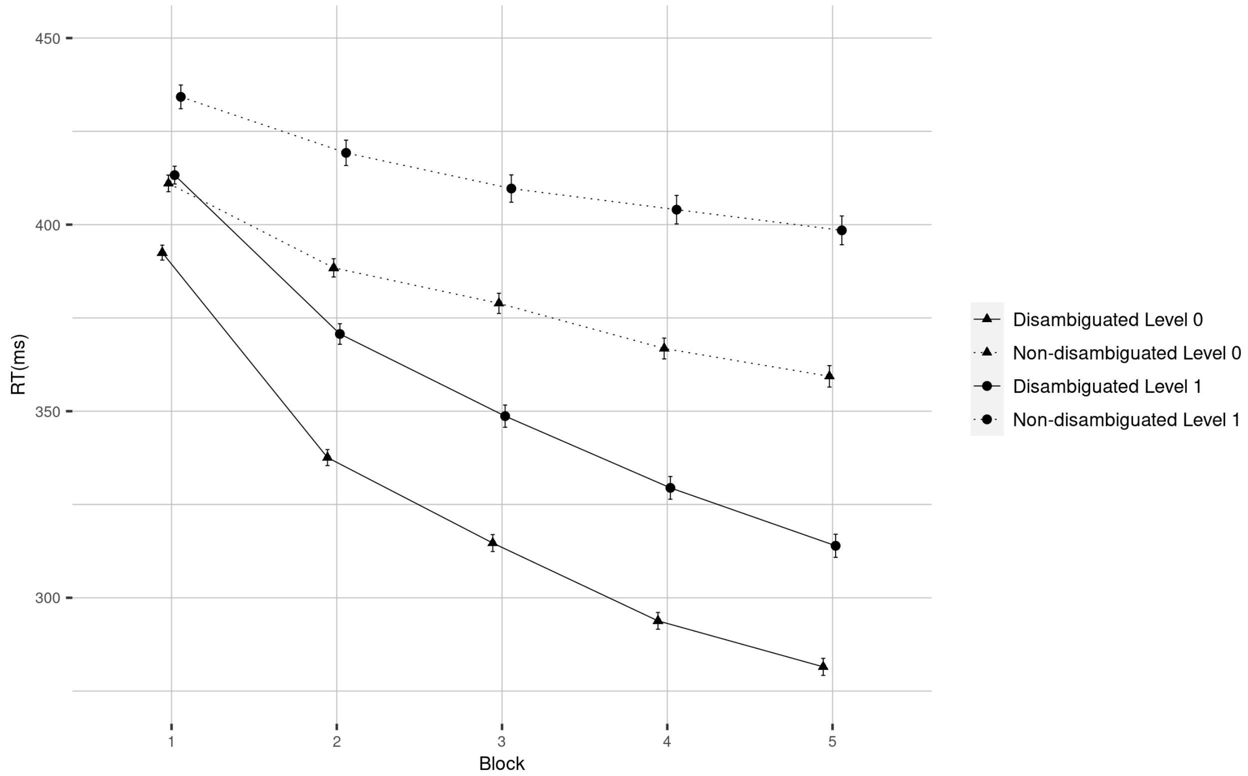


Fig 2. Mean RT (ms) for Disambiguated and Non-disambiguated points of Hierarchical Levels 0 and 1 by Block. Errors bars denote the 95% confidence interval.

Table 1

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Disambiguated and Non-Disambiguated Points by Hierarchical Levels and Blocks

		Block 1		Block 2		Block 3		Block 4		Block 5	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 0	Disambiguated	0.98	0.13	0.99	0.09	0.99	0.11	0.99	0.10	0.99	0.09
	Non-disambiguated	0.93	0.26	0.91	0.28	0.90	0.30	0.89	0.31	0.89	0.31
Level 1	Disambiguated	0.96	0.20	0.96	0.19	0.96	0.19	0.97	0.18	0.97	0.17
	Non-disambiguated	0.91	0.28	0.89	0.32	0.87	0.34	0.86	0.35	0.86	0.35
Level 2	Disambiguated	0.92	0.27	0.91	0.28	0.9	0.30	0.9	0.30	0.9	0.24
	Non-disambiguated	0.94	0.23	0.90	0.30	0.89	0.31	0.88	0.32	0.88	0.33
Level 3	Disambiguated	0.91	0.28	0.89	0.32	0.87	0.33	0.87	0.34	0.87	0.34
	Non-disambiguated	0.91	0.28	0.89	0.32	0.88	0.34	0.84	0.36	0.85	0.36

2.3.1.2 Processing of hierarchical regularities (Levels 1-4)

Hierarchical processing at level 1

Analyses of reaction times showed a main effect of *Exposure* (β = -18.39, *SE* = 0.31, *t* = -59.45, *p* < .001) with a mean reduction of reaction times of 73 ms from block 1 to block 5. There was also a main effect of *Ambiguity level*$_1$ (β = -56.19, *SE* = 0.92, *t* = -61.31, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 56 ms. The interaction *Ambiguity level*$_1$* *Exposure* was also significant (β = -15.20, *SE* = 0.64, *t* = -23.62, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block1 - block5}$ = -95 ms) than non-disambiguated points ($M_{block1 - block5}$ = -34 ms). Results are shown in Fig. 2.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.055, *SE* = 0.01, *z* = -5.022, *p*< .001) with a mean reduction of accuracy of 1.2 % from block 1 to block 5. There was also a main effect of *Ambiguity level*$_1$ (β = 1.35, *SE* = 0.03, *z* = 41.903, *p*< .001) with accuracy higher for disambiguated points (*M* = 0.96) than for non-disambiguated points (*M* = 0.88). The effect of *Exposure* significantly interacted with *Ambiguity level*$_1$ (β = 0.20, *SE* = 0.02, *z* = 8.759, *p*< .001) with accuracy increasing for disambiguated points over exposure ($M_{block1 - block5}$ = 0.01) and decreasing for non-disambiguated points ($M_{block1 - block5}$ = -0.05). Results are shown in Table 1.

Hierarchical processing at level 2

Analyses of reaction times showed a main effect of *Exposure* (β = -12.32, *SE* = 0.36, *t* = -34.036, *p* < .001) with a mean reduction of reaction times of 49 ms from block 1 to block 5. There was also a main effect of *Ambiguity level*$_2$ (β = -8.10, *SE* = 1.05, *t* = -7.693, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 8 ms. The interaction *Ambiguity level*$_2$* *Exposure* was also significant (β = -7.75, *SE* = 0.74, *t* = -10.44, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block1 - block5}$ = -61 ms) than non-disambiguated points ($M_{block1 - block5}$ = -31 ms). Results are shown in Fig. 3.

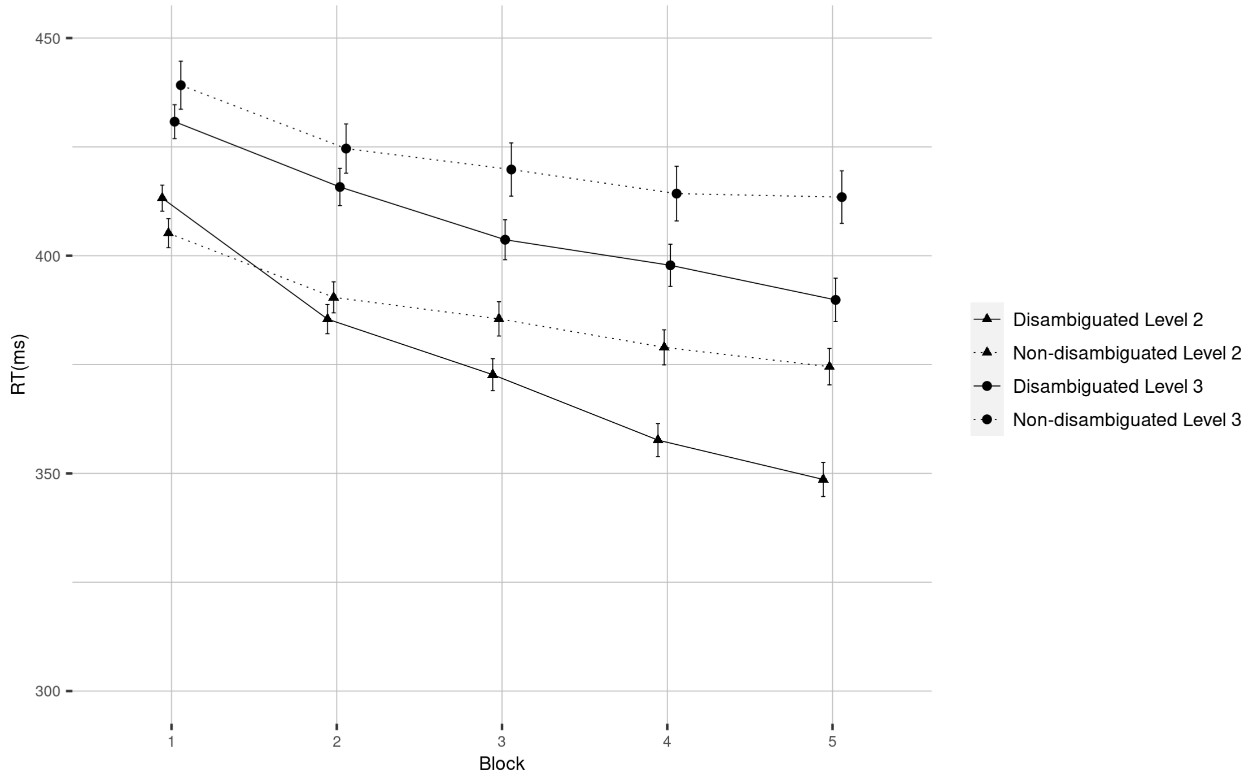


Fig 3. *Mean RT (ms) for Disambiguated and Non-disambiguated points of Hierarchical Levels 2 and 3 by Block. Errors bars denote the 95% confidence interval.*

Concerning accuracy, we found a main effect of *Exposure* (β = -0.10, *SE* = 0.01, *z* = -8.905, *p*< .001) with a mean reduction of accuracy of 3.5 % from block 1 to block 5. There was also a main effect of *Ambiguity* $level_2$ (β = 0.07, *SE* = 0.03, *z* = 2.195, *p* = .028) with accuracy higher for disambiguated points (*M* = 0.91) than for non-disambiguated points (*M* = 0.89). The effect of *Exposure* significantly interacted with *Ambiguity* $level_2$ (β = 0.09, *SE* = 0.02, *z* = 3.811, *p*< .001) with accuracy decreasing less for disambiguated points over exposure ($M_{block1 - block5}$ = -0.02) than for non-disambiguated points ($M_{block1 - block5}$ = -0.06). Results are shown in Table 1.

Hierarchical processing at level 3

Analyses of reaction times showed a main effect of *Exposure* (β = -8.60, *SE* = 0.50, *t* = -17.314, *p* < .001) with a mean reduction of reaction times of 34 ms from block 1 to block 5. There was also a main effect of *Ambiguity* $level_3$ (β = -12.86, *SE* = 1.47, *t* = -8.769, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 12 ms. The interaction *Ambiguity* $level_3$** *Exposure* was also significant (β = -3.224, *SE* = 1.03, *t* = -3.120, *p* = .002) with a more important reduction over exposure for disambiguated points ($M_{block1 - block5}$ = -38 ms) than non-disambiguated points ($M_{block1 - block5}$ = -27 ms). Results are shown in Fig. 3.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.13, *SE* = 0.01, *z* = -9.215, $p < .001$) with a mean reduction of accuracy of 5.2 % from block 1 to block 5. There was also a main effect of *Ambiguity level*$_3$ (β = 0.09, *SE* = 0.04, *z* = 2.187, *p* = .029) with accuracy higher for disambiguated points (*M* = 0.88) than for non-disambiguated points (*M* = 0.87). The interaction *Exposure** *Ambiguity level*$_3$ did not reach significance level (β = 0.05, *SE* = 0.03, *z* = 1.651, $p < .098$). Results are shown in Table 1.

Hierarchical processing at level 4

Analyses of reaction times showed a main effect of *Exposure* (β = -7.46, *SE* = 0.54, *t* = -13.911, $p < .001$) with a mean reduction of reaction times of 30 ms from block 1 to block 5. There was no main effect of *Ambiguity level*$_4$ (β = 0.03, *SE* = 1.56, *t* = 0.023, *p* = .981) and the interaction *Ambiguity level*$_4$* *Exposure* was also not significant (β = 1.59, *SE* = 1.10, *t* = 1.442, *p* = .149).

Concerning accuracy, we found a main effect of *Exposure* (β = -0.16, *SE* = 0.02, *z* = -8.617, $p < .001$) with a mean reduction of accuracy of 6 % from block 1 to block 5. There was no main effect of *Ambiguity level*$_4$ (β = -0.02, *SE* = 0.05, *z* = -0.295, *p* = .768) and the interaction *Ambiguity level*$_4$* *Exposure* was also not significant (β = 0.023, *SE* = 0.04, *z* = 0.603, *p* = .546).

2.3.2. Processing of hierarchical constituency

The results above suggest that participants were sensitive to the higher-order regularities of the sequence up to the third level, we thus restricted the analysis of the structure constituency to level 1, 2 and 3.

Hierarchical constituency at level 1

Analyses of reaction times showed a main effect of *Exposure* (β = -26.52, *SE* = 0.31, *t* = -85.703, $p < .001$) with a mean reduction of reaction times of 106 ms from block 1 to block 5. There was also a main effect of *Structural context*$_{level1}$ (β = 4.89, *SE* = 0.90, *t* = 5.416, $p < .001$) with points in an ambiguous structural context faster than points in a non-ambiguous structural context by 4.9 ms. The interaction *Structural context*$_{level1}$* *Exposure* was not significant (β = -0.86, *SE* = 0.64 t = -1.345, p = .178).

Concerning accuracy, we found a main effect of *Exposure* (β = 0.13, *SE* = 0.03, z = 5.192, p < .001) with accuracy increasing of 0.7 % from block 1 to block 5. There was no main effect of *Structural context*$_{level1}$ (β = -0.02, *SE* = 0.07, z = -0.280, p = .779). However, the interaction *Structural context*$_{level1}$* *Exposure* was significant (β = 0.14, *SE* = 0.05, z = 2.649, p = .008) with accuracy increasing more for points in non-ambiguous structural context ($M_{block1 - block5}$ = 0.009) than points in ambiguous structural context ($M_{block1 - block5}$ = 0.004). Results are shown in Table 2.

Hierarchical constituency at level 2

Analyses of reaction times showed a main effect of *Exposure* (β = -25.32, *SE* = 0.30, t = -83.536, p < .001) with a mean reduction of reaction times of 101 ms from block 1 to block 5. There was no effect of *Structural context*$_{level2}$ (β = -1.29, *SE* = 0.88, t = -1.464, p = .143). The interaction *Structural context*$_{level2}$* *Exposure* was also not significant (β = -0.18, *SE* = 0.58, t = -0.311, p = .756).

Concerning accuracy, we found a main effect of *Exposure* (β = 0.002, *SE* = 0.0004, t = 4.802, p< .001) with accuracy increasing of 0.8 % from block 1 to block 5. *Structural context*$_{level2}$ did not reach significance level (β = -0.002, *SE* = 0.001, t = -1.703, p = .088) and the interaction *Structural context*$_{level2}$ * *Exposure* was also not significant (β = 0.0008, *SE* = 0.0008, t = 0.930, p = .352). Results are shown in Table 2.

Hierarchical constituency at level 3

Analyses of reaction times showed a main effect of *Exposure* (β = -23.18, *SE* = 0.33, t = -68.782, p < .001) with a mean reduction of reaction times of 92 ms from block 1 to block 5. There was also a main effect of *Structural context*$_{level3}$ (β = -4.01, *SE* = 0.98, t = -4.08, p < *.001*) with points in non-ambiguous structural context faster than points in ambiguous structural context by 4 ms. The interaction *Structural context*$_{level3}$* *Exposure* was significant (β = -1.58, *SE* = 0.69 t = -2.279, p = .022) with a more important reduction over exposure for points in a non-ambiguous structural context ($M_{block1\,-\,block5}$ = -94 ms) than for points in an ambiguous structural context ($M_{block1\,-\,block5}$ = -88ms). Fig. 4 shows the results plotted for each disambiguated point of the ambiguous and non-ambiguous structural context.

With respect to accuracy, we found no main effect of *Exposure* (β = 0.0003, *SE* = 0.0005, t = -0.740, p= .459). There was a significant main effect of *Structural context*$_{level3}$ (β = 0.003, *SE* = 0.001, t = 2.222, p = .026) with accuracy better for points in a non-ambiguous structural context (M = 0.96) than for points in an ambiguous structural context (M = 0.95). The interaction *Structural context*$_{level3}$ * *Exposure* was significant (β = 0.002, *SE* = 0.001, t = 2.371, p = .018) with accuracy increasing for points in a non-ambiguous structural context over exposure ($M_{block1\,-\,block5}$ = 0.004) and decreasing for points in an ambiguous structural context ($M_{block1\,-\,block5}$ = -0.005). Results are shown in Table 2.

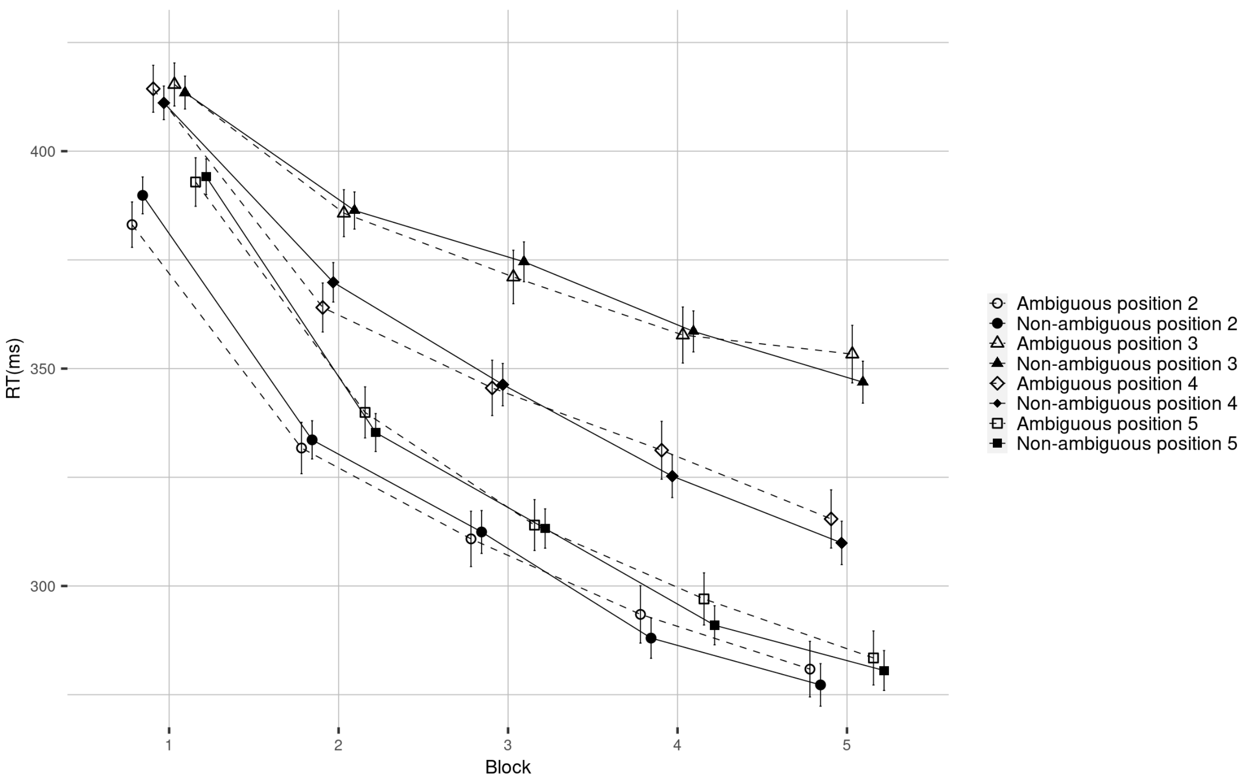


Fig 4. *Mean RT (ms) of disambiguated points occurring in Ambiguous (dashed lines) and Non-ambiguous (solid lines) Structural Contexts at Level 3 by Position and Blocks. The position number indicates the serial order in the constituent [01101], from left to right. Errors bars denote the 95% confidence interval.*

Table 2

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Ambiguous and Non-ambiguous Structural Context by Hierarchical Levels and Blocks

		Block 1		Block 2		Block 3		Block 4		Block 5	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 1	Non-ambiguous Structural Context	0.98	0.13	0.99	0.10	0.99	0.09	0.99	0.09	0.99	0.09
	Ambiguous Structural Context	0.99	0.11	0.99	0.09	0.99	0.11	0.99	0.10	0.99	0.10
Level 2	Non-ambiguous Structural Context	0.97	0.13	0.97	0.12	0.97	0.12	0.98	0.12	0.98	0.12
	Ambiguous Structural Context	0.97	0.12	0.97	0.11	0.97	0.11	0.98	0.11	0.98	0.11
Level 3	Non-ambiguous Structural Context	0.95	0.11	0.96	0.11	0.96	0.12	0.96	0.11	0.96	0.11
	Ambiguous Structural Context	0.95	0.11	0.96	0.11	0.96	0.11	0.95	0.12	0.95	0.12

2.4 Discussion

The aim of the present study was to evaluate if binary sequences can be processed as nested structures. To do so, we created aperiodic self-similar sequences from the Fibonacci grammar, and tested adult participants' learning of their properties in an SRT task. The transitions within these sequences can be considered from a hierarchical point of view. Sequences being self-similar, transitions between units at level n are identical to transitions between constituents at level n+1. At each level, the transitions are either probabilistic or deterministic. Crucially, the probabilistic transitions at level n are embedded in deterministic transitions at level n+1. It is thus possible to reduce the number of probabilistic transitions by recursively embedding deterministic transitions. This recursive structure allows us to predict precisely which unit can be anticipated if the underlying hierarchical structure of the sequence is processed.

We hypothesized that hierarchical processing would result in a progressive construction of the underlying, nested structure. This should be reflected by (a) a progressive ability to anticipate specific points in the sequence that are ambiguous at level n, but disambiguated at level n+1, and (b) a better anticipation for disambiguated points appearing at level n+1 in a constituent following a deterministic transition (non-ambiguous structural context) compared to the same disambiguated points occurring at level n+1 in a constituent following a probabilistic transition (ambiguous structural context).

In line with the first prediction, we found that for levels 0, 1, 2 and 3, disambiguated points showed a steeper reduction of RTs through exposure than their non-disambiguated counterparts. At levels 0 and 1, we also found that through exposure, accuracy increased for disambiguated points while it decreased for non-disambiguated points. However, at levels 2 and 3 accuracy decreased through exposure for both disambiguated and non-disambiguated points suggesting a speed accuracy trade-off. Critically, this decrease in accuracy does not invalidate our predictions since, at level 2, it was significantly greater for non-disambiguated than disambiguated points and at level 3, accuracy was overall higher for disambiguated points. The decrease in accuracy could be due to the boredom of the participants caused by the simplicity of the task. It could also be due to the instructions which only concerned the speed of response. It should also be noted that the magnitude of this decrease remains relatively small, it was at most at 6% between the first and the last block of the experiment. Finally, we found no sign of anticipation at level 4. Taken all together the results of the first analysis suggest that participants were able to build the structure up to the third hierarchical level.

An alternative explanation based on linear precedence may account for the better anticipation of disambiguated points compared to non-disambiguated points. This explanation is based on the fact that disambiguated points are systematically preceded by a specific sub-sequence that never precedes non-disambiguated points of the same level, whereas transitions between sub-sequences of identical length and their following non-disambiguated points are probabilistic (Fig. 1E). Thus, the better anticipation of disambiguated points can potentially come from their linear precedence. However, accounting for the anticipation of disambiguated points with linear precedence faces numerous challenges. First, such explanations would be very costly in terms of memory resource. The linear sub-sequences needed to anticipate the disambiguated points overlaps (see Fig. 5), hence the parser would need to track in parallel all the different patterns. Second, the sequence being binary, the patterns are distinguishable only by their positional order; the parser must therefore also be able to deal with the interference caused by the similarity in the patterns' elements. Finally, the pattern allowing anticipation of disambiguated points would have to be held in memory for a relatively long time. In the present experiment, the pattern retention time would include the 500 ms of the response-to-stimulus interval and the time to answer the trial. If we consider a mean reaction time of 300 ms per trials, the patterns allowing anticipation of disambiguated points at level 1, 2 and 3 should be held in memory for 1.6, 3.2 and 5.5 seconds, respectively. Thus, in order to account for the results, a linear precedence parser would have to overcome these three requirements: overlapping patterns, interference caused by item similarity and long retention time in working memory. The attentional cost induced by these constraints' casts doubt that a simple pattern recognition mechanism could be a plausible candidate to account for anticipation of disambiguated points.

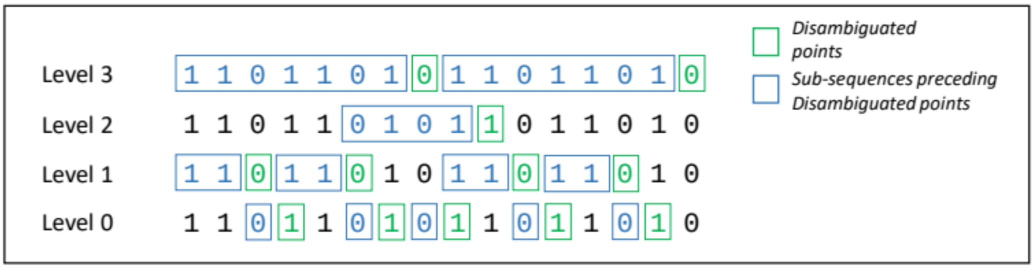


Fig 5. *Sub-sequences (blue) preceding disambiguated points (green) by hierarchical levels. We see that the linear sub-sequences necessary to anticipate the disambiguated points of each levels overlap.*

These results also seem to contradict the hypothesis put forward by Vender et al. (2020) to explain the processing of the Fibonacci grammar. According to this hypothesis, participants would identify certain points of the grammar, called k-points, as relevant structural units and would rely on these k-points to build the hierarchical structure. According to our notation, k-points are the non-disambiguated points of level 0 (i.e., they are all the 1s that appear after [01]) and therefore, level 2 contrast different instance of k-points (i.e., the disambiguated and non-disambiguated points of level 2 are all and only k-points). If the formal status of k-points was at the origin of the building of the hierarchical structure, then they should all be identified in the same way, which should translate into an identical processing advantage for all k-points. Therefore, the difference between disambiguated and non-disambiguated points we found at level 2 cannot be explain by Vender et al. (2020) hypothesis. Moreover, since k-points are by definition 1s, this hypothesis cannot explain the effects we found at level 3, which concern differences between 0s. Since Vender et al's (2020) core argument in favor of k-points relies on the comparison between the processing of k-points in the Fibonacci grammar and in an alternative grammar, we cannot assess the validity of this hypothesis, however, the formal approach adopted by these authors needs to be further elaborated to account for our results.

Concerning the second prediction, we found that accuracy increased significantly more in non-ambiguous structural contexts than in ambiguous structural contexts at level 1[3], suggesting progressive learning of the constituent structure at level 1. Results also showed that points occurring in an ambiguous structural context were overall faster than when they appeared in a non-ambiguous structural context. However, that effect was there from the beginning of the sequence, i.e., it did not interact with exposure, which suggests that it does not reflect learning. Level 3 showed the predicted effect of structural context in both RTs and accuracy, with a significant reduction of RTs and a significant accuracy increase for the non-ambiguous structural context compared to the ambiguous structural context. However, at level 2, we found no effect of structural context in either RTs or accuracy, although a trend was found in the expected direction. Before reasoning about the possible explanation to the lack of effect at level 2, it is important to highlight that the effects found at levels 1 and 3 already exclude the possibility that performance is *only* due to "flat" statistical learning processes (i.e. linear precedence). If better anticipation for the disambiguated points was due to participants memorizing the sub-sequence preceding them, the structural context in which they occur should have no influence given that in both ambiguous and non-ambiguous structural contexts, disambiguated points were preceded by exactly the same sub-sequences. These effects can

3 The reader may find it surprising that accuracy increases with exposure at levels 1 and 2 in the analysis *Processing of hierarchical constituency* while it decreases for the same levels in the analysis *Processing of hierarchical structure*. This is explained by the fact that the two modalities of the factor *Structural context* contrast different instances of disambiguated points, non-disambiguated points are not taken into account. *Structural context*$_{level1}$ contrasts disambiguated points at level 0 and *Structural context*$_{level2}$ contrasts disambiguated points at level 0 combined with disambiguated points at level 1. Since accuracy increased with exposure for the disambiguated points at levels 0 and 1 (see Table 1), it is logical that it also increases for the structural context at levels 1 and 2.

only be accounted for by a strategy that incorporates in one way or another the notion of hierarchy. But why did structural context fail to significantly affect performance at level 2? Although we are currently unable to provide one fully satisfying explanation, we can sketch different lines of reasoning. First, it should be kept in mind that for the analysis of structural context, we compared at each level different instances of the same disambiguated points. At level 2, we compared two subsets of disambiguated points from levels 0 and 1 whose transitional probabilities were $p(1|0) = 1$ and $p(0|11) = 1$, respectively. It could be that the linear precedence of the points involved in this comparison has hidden the effects of structural context. In line with this interpretation, the first analysis showed that these disambiguated points were learned very early in the experiment, already in block 1 (see Fig. 2). Moreover, these disambiguated points were the ones that showed the highest RT decrease. It is thus possible that a floor level was reached, making the effect of structural context undetectable. However, according to this interpretation, the effects should be weaker for the lower level than for higher levels (i.e., it should be the strongest at level 3, followed by level 2 and then level 1) because higher level constituents contain points that are also disambiguated at higher levels, which imply that the influence of the linear precedence should decreases the higher one progresses in the hierarchy. The fact that we observed an effect at level 1 therefore tempers this interpretation, although the effect size was small. Finally, Fig. 4 shows that the effect of structural context at level 3 is distributed across all the points of the constituent. In particular, the RTs of the points at positions 4 and 5, which correspond respectively to disambiguated points at level 1 and 0, decrease more strongly in the non-ambiguous structural context than in the ambiguous structural context. These points are precisely the disambiguated points taken in the analysis of the structural context of level 2. Thus, it might be that the null result found at that level was due to a lack of statistical power.

Taken together, those results suggest that participants have organize the input in a hierarchical way. However, the exact nature of the representations that have been acquired remains to be explored. Fig. 4. shows that the advantage for the non-ambiguous structural context was not driven by one particular point but was distributed across all the points that appeared in that context. This last finding is interesting as it tells us something about the type of hierarchical structure participants built. We have suggested that the process by which participants anticipate higher-order regularities would consist in the recursive combination of units linked through deterministic transitions. However, such a mechanism does not necessarily need to represent a unit as embedded in multiple hierarchical levels; the parser could only retain a representation of the highest level's constituents and anticipate the constituents as wholes. In that view, lower levels' constituents are dissolved into higher levels' constituents and become inaccessible once these higher levels' constituents are represented. In other words, the internal hierarchical structure of the constituents might dissolve as hierarchical building progresses. Such hypothesis is assumed in different models of chunking in which there is no record of the sequential steps by which a chunk is formed (French et al., 2011; Goldwater et al., 2009; McCauley & Christiansen, 2014; Perruchet & Vinter, 1998; Robinet et al., 2011). For example, in PARSER (Perruchet & Vinter, 1998), the system chunks together units that are present in the focus of attention. The span of this focus changes randomly at each trial (encompassing 1, 2 or 3 units). Once a chunk is created, it is processed as a single unit in the focus of attention. Thus, if a chunk reoccurs in the signal, it will occupy only one slot in the focus of attention. This allows the model to chunk multiple chunks together if they are present at the same time in the focus of attention. The activation value of a chunk decreases at each trial if it is not in the focus of attention and increases each time the chunk is encountered. When multiple chunks in memory correspond to the signal (that is when the signal could fit with chunks of different sizes) the activation value of the chunk with the best fit increases while the activation value of the chunks with a lower fit decrease. In this way, the small chunks that are created in the early phases of learning have their activation values progressively tend to 0 as bigger chunks that embed them are created. This results in a representation where only the biggest chunks that fit the signal are kept in memory whereas the smaller chunks that allowed the creation of these bigger chunks are progressively erased from memory. In this view, cognitive representations are limited to chunks with no internal hierarchical structure.

Evidence supporting this claim comes from the so-called *sub-unit effect* that shows that sub-units of a chunk are less accessible once a chunk is learned (Fiser & Aslin, 2005; Giroux & Rey, 2009; Orbán et al., 2008; L. Slone & Johnson, 2015; L. K. Slone & Johnson, 2018). In SRT experiments, this manifests as relatively slow RTs for the first unit of a chunk followed by an acceleration for the remaining units (Hunt & Aslin, 2001; Jiménez et al., 2011; Sakai et al., 2003). In our experiment, if participants were processing constituents as single units without internal structure, RTs should progressively diminish through the constituent. This should be especially true for constituents appearing in the non-ambiguous structural contexts at level 3. This constituent (01101) is composed of 5 points and 4 transitions: if it were processed as a single unit, the transition from one point to the next should result in a progressive reduction of RTs, and the transitional pattern should thus be (- - - -) (where "-"corresponds to a diminution of RTs from each unit to the following). In contrast to that prediction, the transitional pattern observed for this constituent in the last two blocks is (- + - -) (where "+"corresponds to an increase of RTs), i.e., there was a strong deceleration at the second transition. Crucially, that deceleration appears precisely at the border between two constituents at the lower level: the internal structure of [01101] is indeed [[01][101]]. The pattern of acceleration/deceleration therefore provides further evidence that participants represent the internal structure of constituent [01101].

In order to make sure that the deceleration at the second transition observed at the group level was not driven by a subset of participants we computed for each participant the direction of the 4 transitions of the constituent in the non-ambiguous structural context at level 3. We ran by-participants comparisons with 4 linear models (one for each transition). The factor *Position* had two modalities (before, after), "before" coded for the points that was before the transition and "after" coded for the point after the transition. Each model had as predictor the factors *Participants,* and the interaction *Participants* Position* (the factor Position was entered only in the interaction term in order to compare the effect of position for the same individual and not across individuals). In order to increase statistical power, we computed transitions for blocks 4 and 5 jointly (see supplementary materials S1 for detailed results). Table 3 shows the number of participants by transition pattern. We see that 78 % of the participants show a deceleration at the second transition, 22% show no variation in RTs, and critically none shows acceleration. This shows that the transition pattern (- + - -) found at the group level is replicated at the individual level, and is therefore not due to a mix of different patterns across participants. We also see that the transitional pattern (- - - -), expected if chunks lost their internal structure, was found in no participants, suggesting that the constituent [01101] was never processed as a single unit. Crucially, 93 % of the slow-downs occurred at the second and third transitions, that is, at the boundary between lower level constituents. This suggests that participants represent several hierarchical levels simultaneously: the pattern reflects the processing of the internal structure [[01][[1][01]]] of the constituent [01101]. This observation brings further support to our hypothesis that sequences are represented as recursive embedding of constituents.

Table 3

Distribution of the statistical effects for the four transitional patterns in Blocks 4 and 5 combined for the constituent [01101] in Non-ambiguous Structural Context.

Transitional pattern				N° of obs.
Transition 1	Transition 2	Transition 3	Transition 4	
-	+	-	-	30
-	+	=	-	32
-	+	-	=	31
-	+	=	=	22
-	=	=	=	20
-	+	-	+	5
-	=	-	+	3
-	=	-	=	3
=	=	=	=	3
-	+	+	-	2
=	=	-	=	2
-	=	+	-	1
-	=	=	-	1
-	=	=	+	1
=	+	-	=	1
=	+	=	=	1
=	=	-	+	1
-	-	-	-	0

Note. The + sign indicate a significant increase in reaction times. The - sign indicate a significant decrease in reaction times. The = sign indicate no significant differences in reaction times. Significant differences were considered at the $p < 0.05$ level.

In the present study, we proposed that the cognitive system would build a hierarchical structure by recursively combining deterministic transitions in the Fibonacci grammar. This mechanism does not require that participants have access to the rewriting rules of the grammar. Because of the Fib-specific self-similarity which makes the transitional probabilities perfectly scale-free, the surface properties (i.e., the transitional probabilities) lead the parser to a structure that is identical to the natural structure of the Fibonacci grammar. We would like to emphasize that this mechanism is only one possible mechanism to account for the results, which does not mean that it is the only possible strategy to build a hierarchical structure from the Fibonacci grammar. However, it seems that our results can only be explained by a single family of strategies: those that are sensitive to hierarchically organized substrings.

Our results also confirm the finding of Planton et al. (2021) that even sequences as simple as binary sequences can be processed hierarchically. Our proposal that the parser relies on the statistical regularities of the signal to access higher-level constituents is also consistent with the results reported by these authors regarding the involvement of statistical learning. Indeed, this component explained a significant part of the variance even in sequences with high Kolmogorov complexity. The idea that the degree of complexity of the input is the factor that will lead the system to recode the information has also been put forward to explain how the system induces rules from a set of exemplars (E. Pothos, 2010; Radulescu et al., 2019, 2021). In particular, Radulescu et al. (2019, 2021) proposed that the recoding of information into a more abstract format depends on the complexity of the signal and the finite encoding capabilities of the cognitive system. The degree of entropy of a signal (i.e., its complexity) depends on the number of items that compose it as well as on the homogeneity of the distribution of these items. The more homogeneous the distribution (i.e., all items have the same probability) and the longer the signal, the higher the entropy is. Radulescu argues that rule induction arises when the entropy level exceeds the encoding capacity of the system. This upper limit of the amount of information that can be sent through the channel per unit of time forces the system to compress the information into a more abstract format in order to reduce the level of entropy. We suggest that the construction of a hierarchical structure can be seen as a way to reduce the entropic state of the parser: uncertainty is reduced as the hierarchical structure of the signal is built, in line with the proposition of Radulescu et al. (2019, 2021). However, the particularity of the Fibonacci grammar is that at each level, the statistical distribution of the constituents is identical, due to the specific flavor of self-similarity of the Fibonacci grammar. An interesting line for future research could be to ask whether and how self-similarity may play a role in the compression of the input, since it is independent of the entropy of the signal. The rich world of L-systems allows such manipulation, that is manipulating the degree of isomorphism of the self-similarity while keeping entropy constant.

Chapter 3. Uncovering hierarchical structure through statistical dependency

3.1 Introduction

The way in which the cognitive system extracts a hierarchical structure from a linear input is a central issue in cognitive science. In order to anticipate the environment in which it evolves, the cognitive system has to infer from the surface properties of the signal its underlying organization. The human brain is surprisingly good at this task and nested structure extraction has been demonstrated in multiple domains (Dehaene et al., 2015, 2022; Fitch, 2014; Kotz et al., 2018; Lewis & Phillips, 2015; Martins, Krause, et al., 2019; Nakai & Sakai, 2014). However, since the hierarchical structure is not present in the signal itself but must be inferred from it, the cognitive system cannot determine a priori if a signal has an underlying hierarchical organization (Levelt, 2020; Uddén et al., 2020). The question is therefore to know which properties of the signal lead the system to organize it hierarchically.

The *Information Premise* (Pothos, 2010) proposes that when the cognitive system has to process new information, it tends to represent it with the least possible uncertainty. The input is therefore recoded in a format that reduces the entropy of the representational state of the system. Several authors have suggested that the degree of input complexity would modulate the intensity of its compression (Planton et al., 2021; Pothos, 2010; Radulescu et al., 2019, 2021). The underlying reasoning is that processing cost increases with signal complexity. The cognitive system having a finite encoding capacity in terms of memory and processing speed, a signal whose complexity would induce costs that exceed these capacities would push the system to compress it into a more abstract format, thus reducing the cost of information processing. In this view, hierarchical structure building can therefore be seen as a way to condense information by getting rid of item-specific properties and keeping only the properties shared among items.

Radulescu et al. (2019) explored how signal complexity affects rule induction by manipulating the entropy of the input. Entropy is a theoretical measure that quantifies the degree of uncertainty (i.e., the complexity) of a signal (Shannon, 1948). It is a function of the length and number of different items that compose the signal (i.e., the more the signal contains different items, the higher the entropy) as well as the homogeneity of the probability distribution of these items within the signal (i.e., entropy increases when all items have the same probability). Based on previous work of Gómez & Gerken, (2000), the authors distinguished two processes of rule induction. The first process, called *item-bound generalization,* is a form of rule induction that refers to the ability to link perceptual features of different items, that is, generalizing a relation between items' physical dimensions. For example, extracting a rule like "ends in *ba*" when presented with a list of pseudowords that all have "*ba*" as last syllable. *Category-based generalization,* on the contrary, is the process by which rules operating on variables are extracted, that is, rules that apply to the relation between abstract categories that can take any value. For instance, extracting the structure AAB from the pseudo-word "*duduba*" (i.e., a repetition of the first syllables followed by an alternation) and generalizing it to novel stimuli like "*pipiro*" (Marcus et al., 1999). Radulescu et al. (2019) explored whether the degree of entropy of the input (i.e., its complexity) favored one or the other rule induction process. To this end, they presented to participants sequences that contained multiple 3-syllable pseudowords that followed the AAB structure. Participants were then asked to perform a grammaticality judgment (yes-no) on pseudowords that also followed the AAB structure and were made of either identical syllables than the ones used in the familiarization or new syllables. Crucially, they manipulated the entropy of the input by varying the number of different syllables in the pseudowords during the familiarization phase. They found that the increase in entropy went hand in hand with an increase in the acceptance rate of AAB pseudowords made of new syllables. Moreover, as entropy increased, the difference in acceptance rate between the AAB pseudowords made of new syllables and the AAB pseudowords made of familiar syllables decreased. Thus, as the input entropy increases, participants discriminate less between the individual AAB pseudowords from the exposure phase and those of the test phase, demonstrating that the increase in complexity of the input pushes the cognitive system towards a category-based generalization.

Using a different measure of complexity than Radulescu et al. (2019), Planton et al. (2021) showed that increasing the complexity of the input causes the cognitive system to compress the signal representation even when there is no “rule” to learn. Participants were first exposed to a short binary sequence. In a subsequent test phase, they were exposed to an altered sequence that deviated by one item from the initial sequence. The task was to report as quickly as possible if the test sequence deviated from the original. The task depended upon the manipulation of the *Kolmogorov complexity* of the training sequences. Kolmogorov complexity is a theoretical measure where the complexity of a sequence is equal to the size of the shortest computer program that can generate it. The more the Kolmogorov complexity of a sequence increases, the more its most compressed representation requires the use of instruction nesting. In order to determine whether participants had a compressed representation of the sequence, the authors also manipulated the *Shannon surprise* induced by the deviant stimulus in the test sequence. Shannon surprise (Shannon, 1948) measures the degree of uncertainty of an observation according to the history of previous observations and reflects statistical learning. This measure is therefore independent of Kolmogorov complexity because it varies according to the position of the deviant in the sequence, whereas Kolmogorov complexity characterizes the sequence as a whole. The authors found that both complexity and surprise were significant predictors of performance, suggesting that alongside statistical learning, participants recoded the sequence into a compressed representation. They interpreted this as suggesting that the participants have at their disposal a recursive compression algorithm that allows them to recode inputs as simple as binary sequences.

Recent research (Schmid et al., 2023a; Vender et al., 2020) suggests that a parameter independent of complexity to which the cognitive system may also be sensitive to is signal self-similarity. These studies have explored hierarchical structure extraction by presenting to participants aperiodic self-similar sequences that were generated by the Fibonacci grammar. Due to self-similarity, the sequence's surface statistical regularities are identical to the regularities between higher order constituents and because the distribution of sequence's units is aperiodic, successful learning cannot be done by the use of low-level strategies like detecting recurring patterns.

A number of studies that have explored structural processing in the Fibonacci grammar (Fib henceforth) actually failed to provide clear-cut evidence in favor of hierarchical processing (Geambaşu et al., 2016, 2020; Vender et al., 2019). Geambaşu et al. (2016), presented sequences from the Fib grammar consisting of two sounds. After an exposure phase, participants did a grammaticality judgment task. Although 5 participants performed above chance level, the study did not show any effect at the group level. One of the explanations given by the authors for this null result is that some of the foils, although never heard by the participants, were in fact Fib-grammatical (i.e. they were possible subsequences of the Fib grammar). It is thus possible that this made the discrimination too difficult for the participants. Geambaşu et al. (2020) reused the same paradigm but controlled the structure of the foils more carefully. Participants were then able to correctly classify grammatical and ungrammatical strings in the test phase. The authors claimed that this result could not be explained by surface properties of the foils and the grammatical strings because the minimal non-grammatical sub-sequence in the foils was 15 units long (the total length of the foils was 50 units). Because this size exceeds the working memory span, it is unlikely that participants based their judgment on memorized patterns. The authors therefore concluded that participants had developed an implicit sensitivity to the higher-order regularities of the grammar. However, a detailed analysis of the foils showed that only 2 of the 18 used contained a non-grammatical sub-sequence of at least 15 units. The remaining 16 all contained the sub-sequence [01010], which is non-grammatical in Fib. Since this sub-sequence is 5 units long, it is conceivable that participants relied on this difference to reject foils. Thus, the results can also be explained by a low-level strategy consisting in identifying that the sequence [01010] has never been heard during exposure. These two studies illustrate the limitation of the classical paradigm used in AGL. First, the creation of non-grammatical test strings whose rejection cannot be explained by differences in surface properties is extremely difficult. Furthermore, a growing number of studies have challenged the use of reflection-based measures such as two alternative forced choice to assess learning (Arnon, 2020; Christiansen, 2019; Isbilen et al., 2020; Isbilen & Christiansen, 2022). These measures require participants to deliberate over the memorized material, they therefore also capture decision making mechanisms that can vary greatly from one individual to another, adding noise to the results. As a consequence, the sensitivity, reliability and internal consistency of such measures would be relatively low (Arnon, 2020).

Vender et al. (2019) explored hierarchical processing in the Fib grammar via the use of a modified Simon task. This procedure provides a direct measure of the learning performance throughout the whole task rather than at the end of an exposure session. In this task, a blue or red stimulus appeared at each trial on the right or left side of a screen. Participants had to respond to the color of the stimuli irrespective of its location. The order of appearance of the stimuli was determined by the Fib grammar and incongruent trials (a stimulus appearing on the opposite side of the response key) occurred every sixth trial. The sole objective of incongruent trials was to make the task less repetitive for participants. These authors found that 10-year-old children were sensitive to the surface statistical regularities of the sequences. However, they did not explore whether participants detected higher order sequence regularities.

So far, only Schmid et al. (2023a) and Vender et al. (2020) reported results that suggest hierarchical structure extraction from the Fib grammar. However, the theoretical viewpoints adopted by each of these studies differ substantially. Vender et al. (2020) based their rationale on the formal approach developed by Krivochen et al. (2018) who hypothesized that the parser would take advantage of the isomorphism between the surface properties of the Fib grammar and its structural properties in order to extract a hierarchical structure. The idea being that the detection of statistical regularities would push the parser towards an isomorphic structural hypothesis (i.e., the linear order is symmetric to the hierarchy). We will refer to this hypothesis as the “k-points hypothesis”. Schmid et al. (2023a) adopted a computational approach that attributes less weight to the formal aspects of the Fib grammar. Their idea is that the parser would progressively create a multi-level hierarchical representation by the recursive application of transitional probabilities. This would be made possible by the fact that the statistical distribution of units within the Fib sequences is self-similar. The surface properties of the input would therefore drive the cognitive system to build a hierarchical structure that mimics the natural structure of the Fib grammar. We will refer to this hypothesis as the “recursive merge hypothesis”. The term “recursive merge” is to be interpreted literally, i.e., as the recursive combination of two elements. It does not refer to Chomsky's minimalist program (Chomsky, 1995). Note that although this hypothesis does not rely on the property of isomorphism of the Fib grammar emphasized by Schmid et al. (2023a), it does not rule out that it may also play a role; that is, the two hypotheses are not mutually exclusive.

The goal of the present study is to compare the predictions made by each proposal in the processing of an alternative grammar, the Skip grammar, which shares some superficial similarities with the Fibonacci grammar but is substantially different. The details of each of those proposals is described in the following paragraphs, in relation to the specific properties of the Fib and Skip grammars.

The Fib grammar is derived from the Lindenmayer formalism in which there is no distinction between rewriteable and non-rewriteable symbols, and rewrite rules apply simultaneously to all symbols rather than sequentially from left-to-right in a string (Lindenmayer, 1968; Vitányi & Walker, 1978). The Fib grammar shown below consists in two rewriting rules and contains a two-symbol alphabet:

$$0 \rightarrow 1$$

$$1 \rightarrow 01$$

The interpretation of this formalism is the following: [0] is rewritten as [1] and [1] is rewritten as [01]. For the sake of simplicity, we will use the term "point" to refer to the symbols generated by the grammar for the remaining of this article (formally, the rewriting rules of a grammar operate on the "symbols" of an alphabet that differ from their actual realization but this distinction is irrelevant in this case). The successive application of these rules generates longer and longer sequences of points that correspond to different “generations” of the grammar. Each generation corresponds to the concatenation of the two previous ones, therefore the number of points in each generation corresponds to the Fibonacci sequence from which the grammar takes its name (Fig 1C). Despite the simplicity of these rules, the Fib grammar produces a complex structural pattern. The distribution of points within this grammar is asymmetrical: 1s are more frequent than 0s and the ratio of the number of 1s to 0s approximates the golden ratio (1.618). Two transitions are possible in those sequences (from [0] to the next points and from [1] to the next points) and the probability of those transitions is also asymmetric. The transition from [0] to [1] is deterministic: [0] is always followed by [1] $(p(1|0)=1)$. The transition from [1] to the next points is probabilistic: [1] is followed by [0] in 62% of the cases ($p(0|1)= .618$) and by [1] in 38% of the cases ($p(1|1)= .382$). As we mentioned above, the strings produced by this grammar are self-similar and aperiodic. Self-similarity comes from the recursive nature of the rewriting rules. Each generation being the concatenation of the two previous ones, any generation can be segmented into two smaller generations. For example, generation 4 [01101] can be segmented in generations 2 and 3 [[01] [101]] or in generations 1 and 2 [[01][1][01]]. Thus, each generation is a natural constituent of the grammar. The Fib grammar being self-similar, the transitions between points mentioned above are identical to the transitions between constituents (Fig 1A right panel). The aperiodicity is also scale-free: whichever hierarchical level is considered, there is no linear function that can predict with certainty the whole sequence.

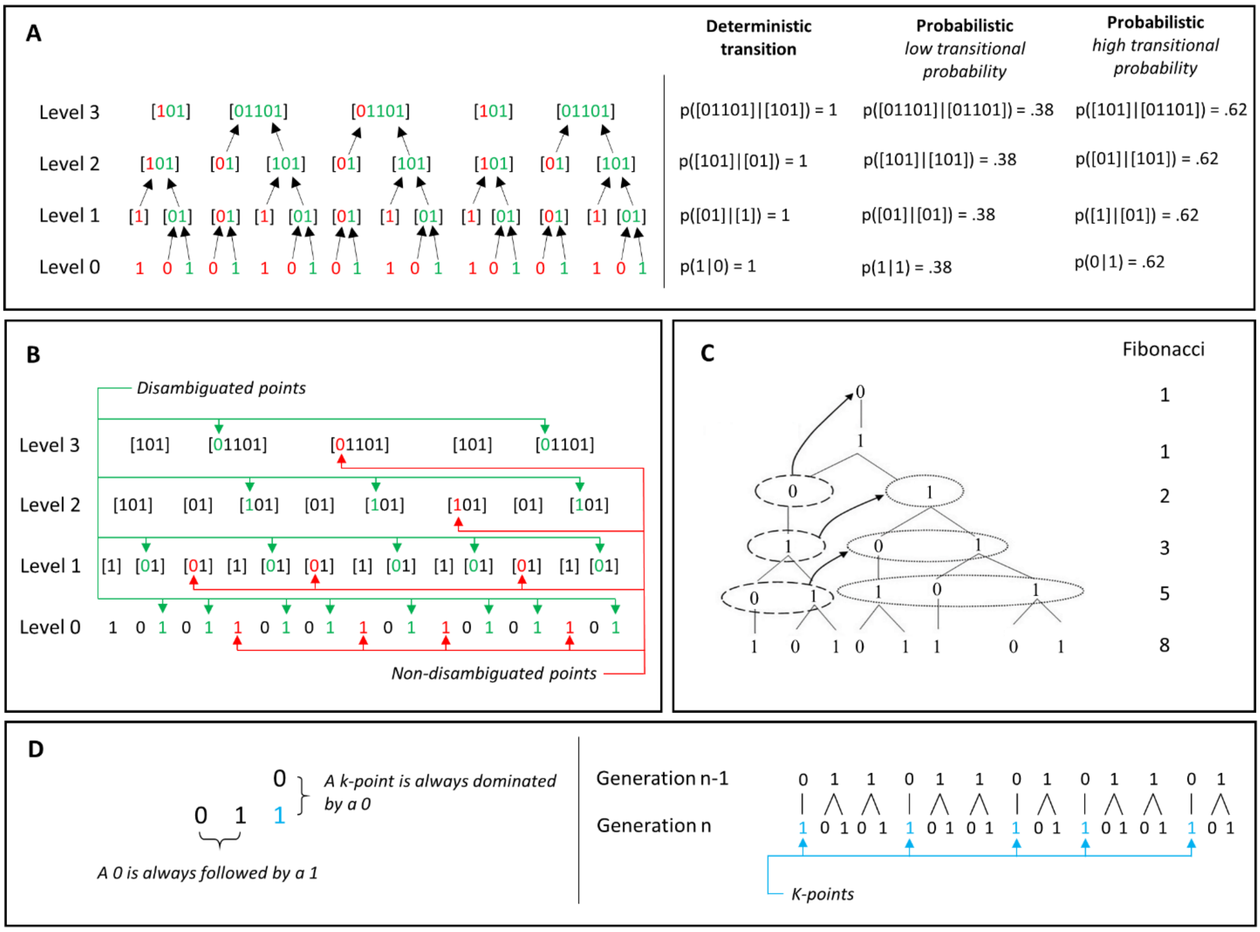


Fig 1. (A) Left panel: depiction of the first three hierarchical levels of generation 7 of the Fib grammar. Points following a probabilistic transition at each level are highlighted in red and points following a deterministic transition in green. To form a new hierarchical level, points/constituents that span across a deterministic transition are combined together (this is illustrated by the arrows). The result is a new representation of the string that consists in the combination of points corresponding to natural higher-order constituents of the grammar (illustrated by the brackets). At each level, constituents spanning a deterministic transition can be combined to form an embedded hierarchy. Right panel: transition probabilities between constituents at each level. (B) Disambiguated points (green) and non-disambiguated points (red) for each hierarchical level for generation 7 of the Fib grammar. In the present study, we used sequences of 120 points. We did not illustrate a full string due space limitation, but the rationale is identical. (C) Derivation of the Fib grammar for the first 5 generations. The right column shows the number of points at each generation, which maps the Fibonacci sequence. Arrows and circles highlight the hierarchical constituency of the grammar. (D) Left panel: depiction of the isomorphism between the linear order “a 0 is always followed by a 1” and derivational order of k-points “a k-point is always dominated by a 0”. Right panel: depiction of k-points at generation n and generation n-1. K-points are highlighted in blue.

The k-points hypothesis proposed by Vender et al. (2020) assumes that some specific points in the grammar, called *k-points*, would allow the construction of the local hierarchical structure of the sequence due to their specific structural status (Fig 1D right panel). Krivochen et al. (2018) defines a k-point as a 1 that is derived from a [0] at generation n-1. Because a k-point is the daughter of a [0] and the sequence 00 is ungrammatical in Fib, the mother of a k-point at generation n-1 is always preceded and followed by a [1], which implies that at generation n, a k-point is always preceded and followed by [01]. This means that the identification of k-points as such gives information about its local environment. The surface expression of k-points corresponds to every [1] appearing after [01] which is the last 1 of the 3-gram [011]. By hypothesizing that the linear order and the structure are symmetrical, the parser could identify k-points because their derivation order "*a k-point is dominated by a 0*" is isomorphic to the linear order of the deterministic transition "*a 0 is always followed by a 1*" (Fig 1D left panel). In fact, identifying k-points as such makes the structure completely transparent because the k-points can be seen as "errors" in the signal: if we remove them, the sequence becomes a periodic sequence of alternating 0s and 1s.

To explore whether the cognitive system is sensitive to the structural status of k-points, Vender et al. (2020) compared the processing of k-points in the Fib grammar to points in an alternative grammar where the surface expression of k-points is present but where they have no special structural status. This alternative grammar, called Skip, is shown below :

$$0 \rightarrow 01$$

$$1 \rightarrow 01101$$

Skip's rewriting rules correspond to two non-consecutive generations of the Fib grammar: [0] is rewritten as Fib generation 1 and [1] is rewritten as Fib generation 3. As a result, Skip displays identical surface properties to Fib: [0] is always followed by [1] ($p(1|0)$ = 1), the sub-sequence [11] is always followed by [0] ($p(0|11)$ = 1) and the first order transitional probabilities are relatively similar : $p(0|1)$ = .73 and $p(1|1)$ = .27. However, in Skip, the linear order is not isomorphic to the derivational order, meaning that although the surface expression of the k-points is present (Skip contains the 3-gram 011), k-points in Skip do not convey structural information as they are daughters of a [1] (Fig 2D). Hence, if k-points were assigned a specific structural status in the processing of Fib-derived sequences, the subsequent presentation of a Skip-derived sequence should not give rise to facilitation in the processing of k-points as they don't have a particular structural status in this grammar.

To put this hypothesis at test, Vender et al. (2020) exposed children to the Fib grammar by using the same Simon task as Vender et al. (2019). After exposition to Fib, the authors added a final block within which the order of stimuli followed the Skip grammar. They found slower processing speed for the k-points in the Skip block compared to the Fib blocks. Thus, participants exhibited a slowdown when they encountered the surface expression of k-points in Skip. Vender et al. (2020) interpreted this result as suggesting that participants were sensitive to the structural status of the Fib k-points.

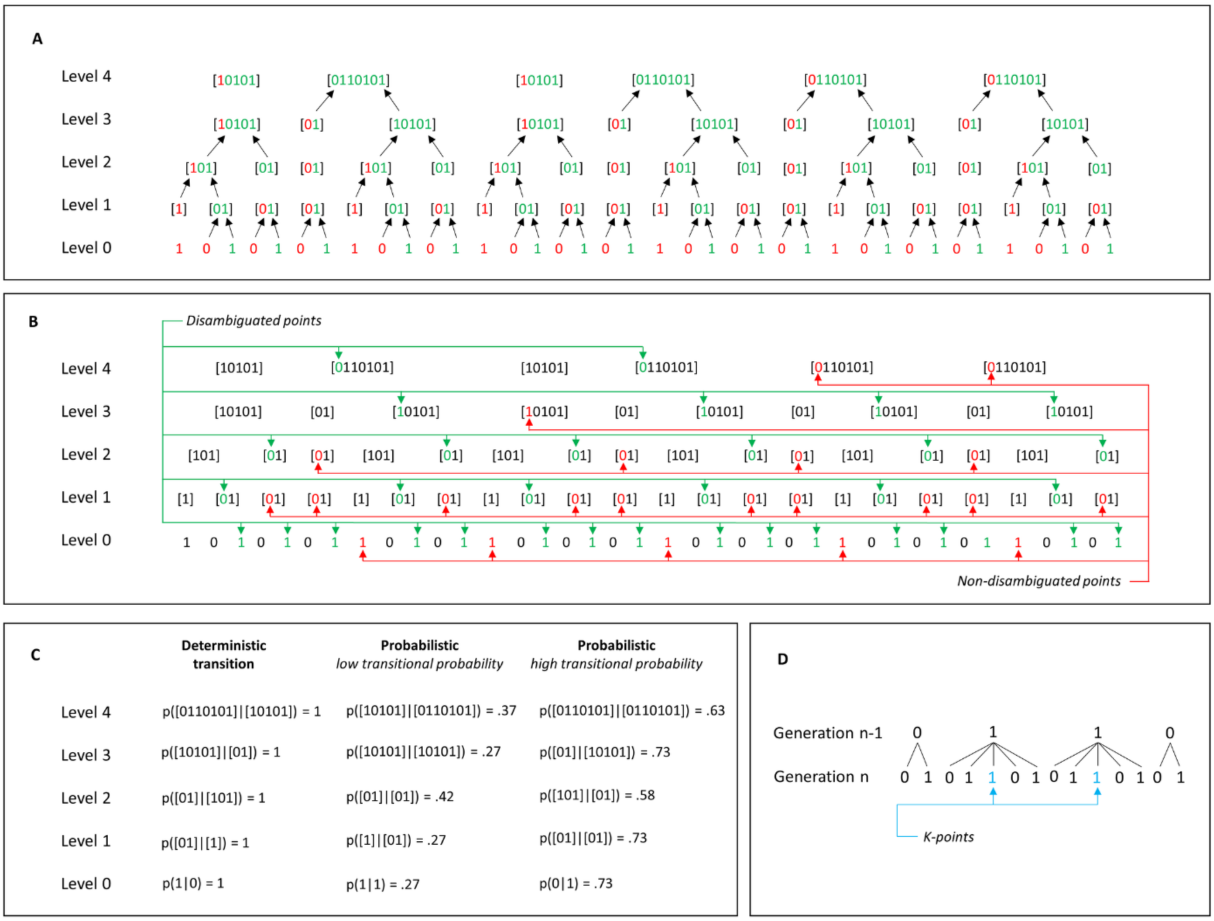


Fig 2. (A) Depiction of the first four hierarchical levels in a string of the Skip grammar. Points following a probabilistic transition at each level are highlighted in red and points following a deterministic transition in green. To form a new hierarchical level, points that span across a deterministic transition are combined together (this is illustrated by the arrows). The result is a new representation of the string that consists in the combination of points corresponding to higher-order constituents of the grammar (illustrated by the brackets). At each level, constituents spanning a deterministic transition can be combined to form an embedded hierarchy. (B) Disambiguated points (green) and non-disambiguated points (red) for each hierarchical level of the Skip grammar. In the present study, we used strings of 120 points. We did not illustrate a full string due space limitation, but the rationale is identical. (C) Transition probabilities between constituents at each hierarchical level of the Skip grammar. (D) Depiction of k-points at generation n and generation n-1 in the Skip grammar. K-points at generation n are highlighted in blue.

However, Schmid et al. (2023a) pointed out that the surface regularities of Skip differ from Fib as soon as the second order conditional probability is considered. In Skip, k-points have a second order conditional probability of $p(1|01)$ = .36 while in Fib it is equal to $p(1|01)$ = .62. Thus, the slower processing observed for the last [1] of the 3-gram [011] in Skip could also be explained by participants becoming sensitive to the fact that [01] is more frequently followed by [0] than by [1] as compared to what they had been exposed to before, i.e., Fib blocks. The effect can therefore also be explained by '*flat*' statistical processing.

Schmid et al. (2023a) studied the extraction of hierarchical structure in the Fib grammar using a similar experimental setup as Vender et al. (2019, 2020), except that the stimuli were presented in the center of the screen. In their proposal, k-points are not assigned any particular status, as hierarchical structure extraction is assumed to take place through the recursive merge of deterministic transitions. At the surface level, points that follow a deterministic transition can be predicted based on the point that precedes them (i.e., a 0 is always followed by a 1) whereas this is not the case for points that follow a probabilistic transition. However, as we explained above, the self-similarity in Fib is such that the surface transitional probabilities are identical to those between constituents of higher hierarchical levels. This implies that constituents following a higher-order deterministic transition can be predicted on the basis of the constituent that precedes them (Fig 1A right panel). Crucially, points that follow a probabilistic transition at level n can appear inside a constituent that follows a deterministic transition at level n+1. Therefore, the detection of higher-order deterministic transitions allows to predict some of the points that follow a probabilistic transition at lower levels. To access this constituent structure, the processing mechanism would start by merging the points linked by a deterministic transition, and then use the output of this process to detect the deterministic transitions at the next hierarchical level. This process of recursive merge would progressively transform the representation of the sequence into a complex hierarchical structure of embedded constituents (Fig 1A left panel). For example, merging the points that span across the deterministic transition at level 0 ($p(1|0) = 1$) leads to the creation of the constituent [01]. Therefore, the sequence can be represented at level 1 with two constituents: the newly created constituent [01] and the constituent [1] (which consists of what remains after the merge). The cognitive system could then detect the higher-order deterministic transition between constituents $p([01]|[1]) = 1$ which would enable to predict a subset of the constituents [01]. Thus, some of the 0s that appear in the constituent [01] (and that always follow a probabilistic transition at level 0) can now be predicted on the basis of the constituent in which they appear. This hypothesis derives precise predictions about which points should be subject to anticipation as each hierarchical level contains points that follow a higher order deterministic transition as well as points that follow a higher order probabilistic transition. The recursive merge hypothesis predicts that the building of the structure would be reflected by the faster processing of points following a deterministic transition at a given level compared to points following a probabilistic transition at the same level (Fig 1B). For clarity, we will use the same terminology as Schmid et al. (2023a) where "*disambiguated points*" refer to points following a deterministic transition at a given level and "*non-disambiguated points*" refer to points following a probabilistic transition at the same level. In line with their hypothesis, results showed that participants were able to anticipate disambiguated points up to the third hierarchical level. They interpreted those results as suggesting that the participants' sequence representation corresponded to the recursive embedding of constituents linked through deterministic transitions.

Interestingly, some aspects of their findings seem incompatible with the k-points hypothesis. Under this hypothesis, all k-points have the same structural status (they are all dominated by a [0] in Fib) and their identification is what allows the building of the hierarchical structure. This hypothesis thus predicts that k-points will be identified in the same way, which should be reflected by an identical processing advantage for all k-points. In contrast to that prediction, Schmid et al. (2023a) found that disambiguated points were processed faster than non-disambiguated points at level 2, despite the fact that at this level, both disambiguated and non-disambiguated points are k-points. Moreover, since k-points are by definition 1s, the hypothesis that structure building operates on the basis of k-points cannot explain the effects they found at level 3, where disambiguated points were processed faster than non-disambiguated points despite the fact that at this level, both disambiguated and non-disambiguated points are 0s (Fig 1B). It thus seems that in its current state, the k-points hypothesis cannot explain the results of Schmid et al. (2023a).

Nevertheless, the core argument put forward by Vender et al. (2020) in favor of the k-points hypothesis was not based on the processing of k-points in the Fib grammar *per se*, but on the fact that their processing differs across grammars, suggesting that participants were sensitive to specific properties in the Fib grammar that are absent in the Skip grammar. Since the study by Vender et al. (2020) contained only one block of Skip and Schmid et al. (2023a) did not explore the processing of the Skip grammar, a deeper exploration of the processing of the Skip grammar seems necessary.

The purpose of the present study is thus to examine the predictions of the k-points hypothesis as well as those of the recursive merge hypothesis in the processing of the Fib and Skip grammars. To this end, we conducted two experiments using the experimental paradigm of Schmid et al. (2023a). In the first experiment, participants were first exposed to five blocks of the Fib grammar and then to one block of the Skip grammar, in the second experiment, participants were first exposed to five blocks of the Skip grammar and then to one block of the Fib grammar.

The k-points hypothesis proposes that in the Fib grammar, the extraction of the hierarchical structure relies on the isomorphism between the linear and derivational orders that allows the identification of k-points as relevant structural units. This hypothesis assumes that the advantage conferred by the isomorphism is restricted to k-points. Four predictions follow from this hypothesis :

(1) Because all k-points have the same formal status in the Fib grammar, they should all be processed equally well. Thus, there should be no difference between disambiguated and non-disambiguated points on the second hierarchical level in the Fib blocks of Experiment 1 since they are all k-points.

(2) In Experiment 1, the switch from Fib to Skip should negatively affect the processing of k-points in Skip because in this grammar, the regularity "*a 0 is always followed by a 1*" does not give access to the k-points dominance relation. This should be reflected by an increase

of RTs and of the error rate for k-points in Skip compared to the last block of the Fib grammar.

(3) In Experiment 2, the switch from Skip to Fib should positively affect the processing of k-points in Fib because the prior processing of Skip allows the acquisition of the surface regularity "*a 0 is always followed by a 1*". Thus, the parser should be already sensitive to the relevant surface regularity when it encounters Fib, which should facilitate the identification of k-points. This should be reflected by a decrease of RTs and error rates for k-points after the grammar switch.

(4) K-points should show a processing advantage in Fib over Skip even in the absence of prior exposure to an alternative grammar. Thus, k-points should be learned better, i.e., showing a stronger decrease of RTs and higher accuracy, in Fib (Experiment 1) than in Skip (Experiment 2).

The recursive merge hypothesis proposes that the construction of the hierarchical structure is done by a recursive combination of deterministic transitions. This hypothesis predicts a processing advantage that is not restricted to k-points, the determining factor being whether a point follows a higher order deterministic transition or not. Like Fib, Skip contains only one deterministic transition at each hierarchical level. As a result, we can determine unambiguously which points should be subject to anticipation according to the recursive merge hypothesis. Since this hypothesis does not take into account the differences in isomorphism between each grammar, it predicts that hierarchical learning should take place in the same way in each grammar. We can therefore derive the following predictions:

(a) In Experiment 1, the processing of strings generated from the Fib grammar should give rise to better learning for disambiguated points at a given level compared to non-disambiguated points at the same level.

(b) In Experiment 2, the processing of strings generated from the Skip grammar should give rise to better learning for disambiguated points at a given level compared to non-disambiguated points at the same level.

In the remaining part of this paper, we first report Experiment 1 and evaluate predictions (1), (2) and (a). We then report Experiment 2 and evaluate predictions (3) and (b). Finally, we report a last analysis where we compared the results of Experiments 1 and 2 in order to evaluate prediction (4). We found that prediction (2) of the k-points hypothesis was in line with the results, replicating the observation reported by Vender et al. (2020). On the other hand, the results were in contradiction with predictions (1), (3) and (4) of the k-points hypothesis. The k-points hypothesis thus seems to be insufficient to fully account for the processing of the Fib and Skip grammars. Concerning the recursive merge hypothesis, our results were in agreement with prediction (a), suggesting hierarchical learning during Fib grammar processing, in line with the observations reported by Schmid et al. (2023a). Concerning prediction (b), results were in line with the recursive merge hypothesis up to level 3, suggesting that Skip grammar is at least partially processed hierarchically. At level 4, however, results were opposite to prediction (b). We propose an ad-hoc explanation of this result in the discussion based on the difference in the form of self-similarity of Fib and Skip.

3.2 Experiment 1 : Switching from Fib to Skip

Experiment 1 aimed at evaluating predictions (1) and (2) of the k-points hypothesis and prediction (a) of the recursive merge hypothesis. Like in Vender et al. (2020), participants were first exposed to sequences generated by the Fib grammar and then to sequences generated by the Skip grammar. Beyond the fact that the current study was conducted on adults, whereas that of Vender et al. (2020) was conducted on children, our study differs on other aspects. First, instead of a Simon task, we used the simpler task of Schmid et al. (2023a) in which stimuli are all presented centrally. Since Schmid et al. (2023a) failed to find any anticipation effect for the disambiguated points of level 1, we increased the number and length of the blocks to increase statistical power. We also set the

Response-to-Stimulus Interval (RSI) at 1000 ms instead of the 500 ms RSI used in Vender et al. (2020) and Schmid et al. (2023a). Finally, in both Vender et al. (2020) and Schmid et al. (2023a), each block corresponded to a full generation of the Fib grammar (generations 10 and 12, respectively) and the blocks were identical across participants. In the present experiment, each participant saw a different block consisting of a 120-long string taken from generation 23 of the Fib grammar in order to exclude the possibility that the effects come from specific properties of a string.

3.2.1 Methods

3.2.1.1 Participants

Fifty-eight students (12 men and 46 women; mean age 21.5 years old) from an introductory psycholinguistics course from the university of Geneva participated in the experiment for course credits. All participants reported normal or corrected-to-normal vision. This study was approved by the ethics commission of the University of Geneva. Informed consent was obtained from each of the participants included in the study.

3.2.1.2 Materials

In the training block, we used a binary sequence of 120 stimuli. The order of stimuli was randomized for each participant. The sequences used in the experimental blocks corresponding to the Fib grammar were created by extracting 290 sub-sequences of 120 point from generation 23 of the Fib grammar. Each of those 290 sub-sequences corresponded to a unique block and were used only once across participants. The sequences used in the experimental blocks corresponding to the Skip grammar were created by extracting 58 sub-sequences of 120 points from generation 8 of the Skip grammar. Each of those 58 sub-sequences corresponded to a unique block and were used only once across participants.

3.2.1.3 Design and procedure

Each trial consisted of a blue or red circle 100 px in diameter presented at the center of the screen. Blue circles correspond to the 1s and the red circles to the 0s of the grammar. The circles disappeared after the response of the participant, or after 1200ms if no response was given. The response-to-stimulus interval lasted 1000 ms. Participants were asked to press as quickly as possible the button corresponding to the color of the circle (X=blue, N=red) on a QWERTZ keyboard. Since the task was extremely simple, instructions did not emphasize the precision of the answer in order to avoid ceiling effects. No information related to the grammar was given. The experiment started with a training block followed by 6 experimental blocks. The training block was intended to accustom the participants to the task; therefore, we did not analyze it as we had no predictions about it. In experimental blocks 1 to 5, the order of the trials was determined by the Fib grammar. In experimental block 6, the order of the trials was determined by the Skip grammar. The experiment was conducted using an HP elite book laptop running on Windows 7 with a 14’’ inch screen and a resolution of 1920*1080 pixels. The computer was placed at approximately 60 cm from the participants. The experiment used MATLAB version 2016b (MATLAB, 2016) programming software. Instructions were displayed on the screen and participants had to click on a button to start the experiment. Participants were told that the experiment contained 7 blocks and that they could take a pause between each block. Before each block, a message was displayed indicating the number of the blocks remaining and participants had to click a button to start. The experiment lasted approximately 20 minutes.

3.2.1.4 Data analyses

In order to evaluate prediction (2) of the k-points hypothesis, we conducted a first analysis (*Processing of k-points*) where we compared the RTs and accuracy of k-points in the last Fib block to that of the Skip block. The k-points hypothesis predicts slower RTs and higher error rates for the Skip k-points compared to the Fib k-points.

In order to evaluate prediction (a), we analyzed the Fib blocks data (*Processing of Fib hierarchical structure*) as Schmid et al. (2023a). We compared within each hierarchical level the RTs and accuracy of the points that follow a deterministic transition (i.e., disambiguated points) to those that follow a probabilistic transition (i.e., non-disambiguated points) (Fig 1B). The Fib grammar being asymmetric in its distribution of 0s and 1s, we compared within each level only 1s to 1s and 0s to 0s. The recursive merge hypothesis predicts that hierarchical building should result in a larger decrease in RTs and error rates across blocks for disambiguated points compared to non-disambiguated points. The same analysis also allowed us to test prediction (1) that k-points in Fib should all be processed in the same way. If this prediction is correct, there should be no difference between disambiguated and non-disambiguated points at the second hierarchical level since at this level, both types of points are k-points.

We removed five participants who had an error rate superior to 3 *SD* to the mean error rate in at least one block. RTs and accuracy were both modeled as dependent variables. We removed from the analysis all the trials where participants did not respond after 1200 ms (47 trials). For the analysis of RTs, only trials with a correct answer were included. Homoscedasticity and normality were checked by visual inspection of residual plots. Data from the remaining 53 participants were analyzed with linear mixed-effects models as implemented in the lme4 package for R (Bates et al., 2014; R Core Team, 2022). For all the models, p-values were calculated by way of the Satterthwaites's approximation to degrees of freedom with the lmerTest package (Kuznetsova et al., 2015).

For the analysis *Processing of k-points*, the model included the fixed-effect factor *Grammar*. *Grammar* is a discrete variable contrasting k-points in the last block of Fib (Block 5) and in the block of Skip (block 6). As random effects, the model had an intercept for *Participants.* For the analysis *Processing of Fib hierarchical structure*, models included two fixed-effect factors and their interaction: *Exposure, Fib ambiguity,* and *Exposure*Fib ambiguity*. *Exposure* was treated as a continuous variable with a value of 0 for trials of the 1st experimental block, and of 1, 2, 3 and 4 for trials of the 2nd, 3^{d}, 4th and 5th blocks. Treating this factor as continuous allowed us to have a single estimate that represents the evolution (i.e., the slope) of performance throughout the experiment across all participants. *Fib ambiguity* is a discrete variable contrasting disambiguated and non-disambiguated points and operationalized differently depending on the level at which its effect is explored (it is labeled *Fib ambiguity* $level_n$ according to the level at which it has been operationalized). We entered as fixed effects the factors *Fib ambiguity* $level_n$ (Disambiguated vs. Non-disambiguated), *Exposure*, and the interaction *Exposure*Fib ambiguity*. The modality "Non-disambiguated" of the factor *Fib ambiguity* $level_n$ was always set as the intercept of the models. As random effects, the models had intercept for *Participants*.

3.2.2 Results

3.2.2.1 Processing of k-points

Analyses of RTs showed a significant effect of *Grammar* (β = 14.49, *SE* = 4.6, *t* = 3.149, *p* = .001) with k-points in the last block of the Fib grammar (*M* = 347 ms) faster than the k-points in the Skip grammar (*M* = 361 ms). Concerning accuracy, we found a significant effect of *Grammar* (β = -0.56, *SE* = 0.12, *z* = -4.691, *p* < .001) with higher accuracy for Fib k-points (*M* = 0.89, *SD* = 0.31) than Skip k-points (*M* = 0.83, *SD* = 0.38).

3.2.2.2 Processing of Fib hierarchical structure

Processing of surface statistical regularities (Level 0).

Analyses of RTs showed a main effect of *Exposure* (β = -14.18, *SE* = 0.52, *t* = -27.11, *p* < .001) with a mean reduction of RTs of 71 ms from block 1 to block 5. There was also a main effect of *Fib ambiguity* $level_0$ (β = -46.50, *SE* = 1.54, *t* = -30.27, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 47 ms. The interaction *Exposure*Fib ambiguity* $level_0$ was also significant (β = -13.46, *SE* = 1.08, *t* = -12.445, *p* < .001) with a more important reduction over time for disambiguated points ($M_{block5 - block1}$ = -79 ms) than non-disambiguated points ($M_{block5 - block1}$ = -23 ms) ($M_{block5 - block1}$ indicates the mean difference between blocks 5 and 1). Results are shown in Fig 3.

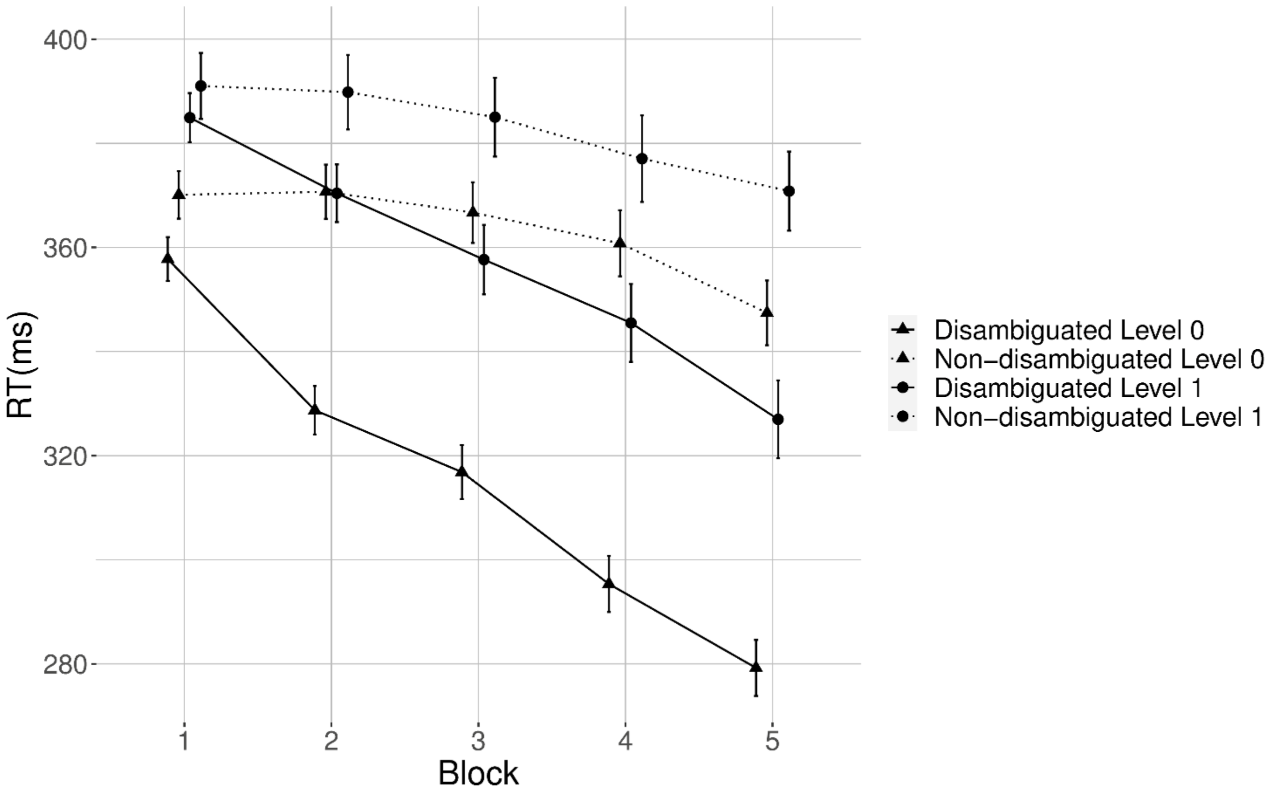


Fig 3. Experiment 1: mean RTs for disambiguated and non-disambiguated points by block (level 0 and 1). Errors bars denote the 95% confidence interval.

Investigating accuracy, we found a main effect of *Exposure* (β = -0.10, *SE* = 0.03, *z* = -3.716, *p* < .001) with a mean reduction of accuracy of 1.4 % from block 1 to block 5. There was also a main effect of *Fib ambiguity level*$_0$ (β = 1.66, *SE* = 0.08, *z* = 20.309, *p* < .001) with higher accuracy for disambiguated points (*M* = 0.98) than for non-disambiguated points (*M* = 0.92). Crucially, the effect of *Exposure* significantly interacted with *Fib ambiguity level*$_0$ (β = 0.31, *SE* = 0.06, *z* = 5.358, *p* < .001) with accuracy increasing for disambiguated points over exposure ($M_{block5 - block1}$ = 0.01) and decreasing for non-disambiguated points ($M_{block5 - block1}$ = -0.05). Results are shown in Table 1.

Table 1

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Disambiguated and Non-Disambiguated Points in the Fib grammar by Hierarchical Levels and Blocks.

		Block 1		Block 2		Block 3		Block 4		Block 5	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 0	Disambiguated	0.98	0.15	0.98	0.14	0.99	0.12	0.98	0.13	0.99	0.12
	Non-disambiguated	0.94	0.23	0.93	0.26	0.93	0.26	0.90	0.30	0.89	0.31
Level 1	Disambiguated	0.95	0.22	0.95	0.21	0.96	0.21	0.94	0.25	0.94	0.24
	Non-disambiguated	0.94	0.24	0.89	0.32	0.87	0.34	0.87	0.34	0.87	0.34
Level 2	Disambiguated	0.93	0.26	0.92	0.27	0.92	0.27	0.89	0.31	0.91	0.29
	Non-disambiguated	0.97	0.18	0.93	0.25	0.93	0.25	0.91	0.29	0.86	0.34

Hierarchical processing at level 1.

Analyses of RTs showed a main effect of *Exposure* (β = -10.59, *SE* = 0.68, t = -15.59, p < .001) with a mean reduction of 53 ms from block 1 to block 5. There was also a main effect of *Fib ambiguity* $level_1$ (β = -24.47, *SE* = 1.99, t = -12.26, p < .001) with disambiguated points being faster than non-disambiguated ones by 24 ms. The interaction *Exposure*Fib ambiguity* $level_1$ was also significant (β = -8.69, *SE* = 1.40, t = -6.19, p < .001) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -58 ms) than non-disambiguated points ($M_{block5 - block1}$ = -20 ms). Results are shown in Fig 3. Concerning accuracy, we found a main effect of *Exposure* (β = -0.12, *SE* = 0.02, z = -4.828, p < .001) with a mean reduction of accuracy of 3.1 % from block 1 to block 5. There was also a main effect of *Fib ambiguity* $level_1$ (β = 0.85, *SE* = 0.07, z = 12.082, p< .001) with higher accuracy for disambiguated points (M = 0.95) than non-disambiguated points (M = 0.89). The interaction *Exposure * Fib ambiguity* $level_1$ was not significant (β = 0.10, *SE* = 0.05, z = 1.923, p = .054). Results are shown in Table 1.

Hierarchical processing at level 2.

Analyses of RTs showed a main effect of *Exposure* (β = -5.55, *SE* = 0.81, *t* = -6.812, *p* < .001) with a mean reduction of 28 ms from block 1 to block 5. There was no effect of *Fib Ambiguity level*$_2$ (β = -4.18, *SE* = 2.37, *t* = -1.768, *p* = .077) . The interaction *Exposure**Fib ambiguity level*$_2$ was however significant (β = -6.546, *SE* = 1.67, *t* = -3.917, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -32 ms) than non-disambiguated points ($M_{block5 - block1}$ = -8 ms). Results are shown in Fig 4. Concerning accuracy, we found a main effect of *Exposure* (β = -0.18, *SE* = 0.03, *z* = -5.909, *p*< .001) with a mean reduction of accuracy of 5.2 % from block 1 to block 5. There was no effect of *Fib ambiguity level*$_2$ (β = -0.08, *SE* = 0.09, *z* = -0.935, *p* = .350). The effect of *Exposure* however significantly interacted with *Fib ambiguity level*$_2$ (β = 0.24, *SE* = 0.07, *z* = 3.618, *p*< .001) with accuracy decreasing less for disambiguated points over time ($M_{block5 - block1}$ = -0.02) than for non-disambiguated points ($M_{block5 - block1}$ = -0.10). Results are shown in Table 1.

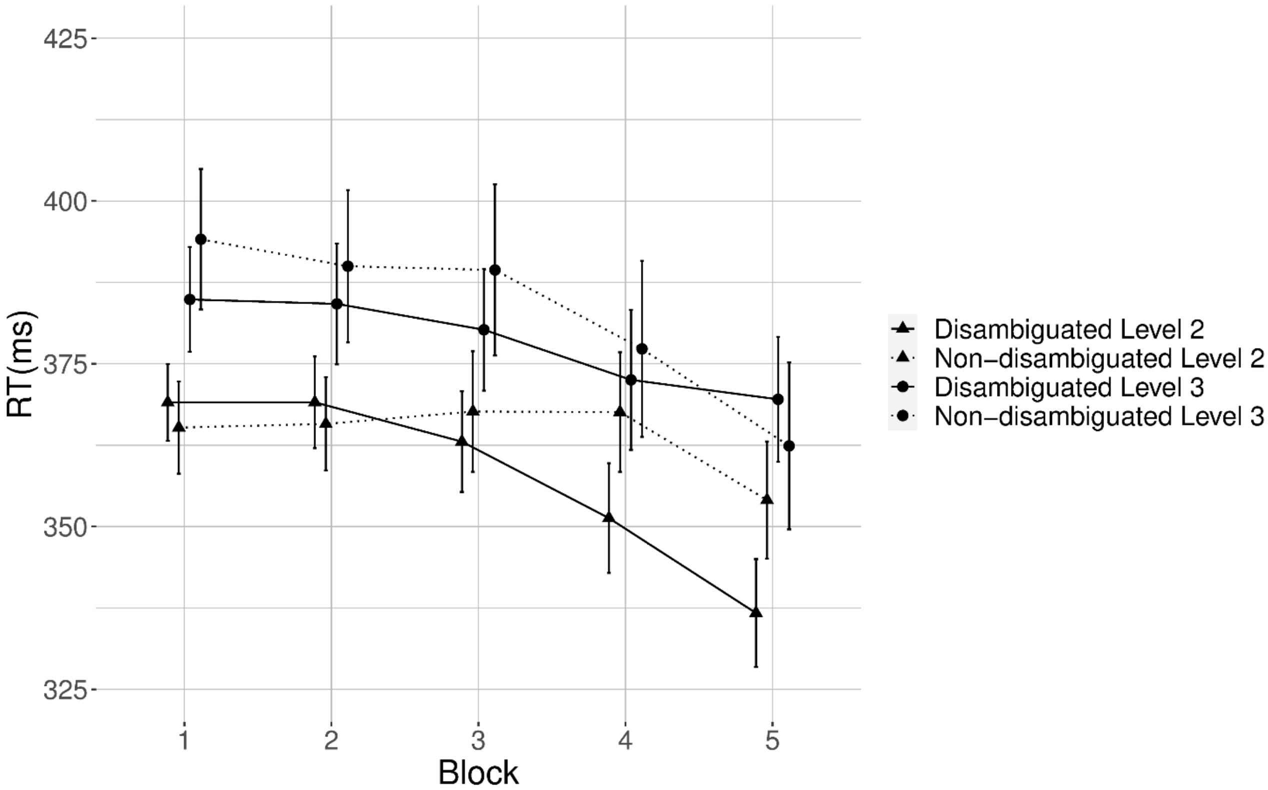


Fig 4. Experiment 1: mean RTs for disambiguated and non-disambiguated points by block (level 2 and 3). Errors bars denote the 95% confidence interval.

Hierarchical processing at level 3.

Analyses of RTs showed a main effect of *Exposure* (β = -5.08, *SE* = 1.06, *t* = -4.788, *p* < .001) with a mean reduction of 26 ms from block 1 to block 5. There was no effect of *Fib ambiguity level$_3$* (β = -3.19, *SE* = 3.11, *t* = -1.024, *p* = .306) and the interaction *Exposure*Fib ambiguity level$_3$* failed to reach significance (β = 3.55, *SE* = 2.18, *t* = 1.626, *p* = .104). Results are shown in Fig 4. Concerning accuracy, we found a main effect of *Exposure* (β = -0.16, *SE* = 0.03, *z* = -4.725, *p*< .001) with a mean reduction of accuracy of 7.1 % from block 1 to block 5. There was no effect of *Fib ambiguity level$_3$* (β = 0.02, *SE* = 0.10, *z* = 0.161, *p* = .872) and the interaction *Exposure*Fib ambiguity level$_3$* was also not significant (β = -0.03, *SE* = 0.07, *z* = 0.488, *p* = .625).

3.2.3 Discussion

The results of Experiment 1 fail to support prediction (1) according to which all k-points should be processed similarly. The effect of ambiguity found at level 2 cannot be explained by the k-points hypothesis: despite the fact that disambiguated and non-disambiguated points were all k-points, they were not processed similarly. However, results validate prediction (2) of the k-points hypothesis, replicating Vender et al. (2020) : when switching from the Fib grammar to the Skip grammar, k-points were processed more slowly and gave rise to more errors than their counterparts in the last block of the Fib grammar.

The results of Experiment 1 also meet prediction (a) of the recursive merge hypothesis: participants learned disambiguated points better than non-disambiguated points at levels 0, 1 and 2. RTs slopes across exposure were steeper for disambiguated points than for non-disambiguated points. Accuracy tended to globally decrease through exposure, which may suggest a speed-accuracy trade-off. However, this does not affect our interpretation of the results since what is key is actually how ambiguity affected accuracy. At levels 0 and 1, accuracy increased for disambiguated points while it decreased for non-disambiguated points (note that at level 1, the interaction just failed to reach significance at *p* = .054) and at level 2, accuracy decreased more strongly for non-disambiguated points than for disambiguated points. These results therefore bring support to the hypothesis that participants built a hierarchical structure up to the second hierarchical level. An overall decrease in accuracy has also been observed in Schmid et al. (2023a), which used the same experimental paradigm. Two non-mutually exclusive reasons have been advanced: the decrease in accuracy could be due to the boredom of the participants caused by the simplicity of the task or to the instructions, which only emphasize response speed.

Nevertheless, in contrast to Schmid et al. (2023a), we failed to find significant structural learning at level 3. Three non-exclusive possibilities can explain this finding. First, the response-to-stimulus interval in the present experiment was 1000 ms instead of 500 ms in Schmid et al. (2023a). Several SRT task studies have reported a decrease in sequence learning with increasing RSI duration (Frensch & Miner, 1994; Soetens et al., 2004; Willingham et al., 1997). This reduced learning was attributed to the reduction of the activation level of the relevant representations in working memory. The longer interval of our experiment could have impaired the hierarchical building process as items had to be kept longer in memory to be merged into constituents, and may have been subject to decay (Barrouillet et al., 2004; Frensch & Miner, 1994; Hommel, 1994; S. T. Mueller et al., 2003), but see Lewandowsky & Oberauer, (2009); Oberauer & Lewandowsky, (2014) for a discussion on decay. Second, the lack of exposition to the grammar might have also played a role. In Schmid et al. (2023a), participants were exposed to 1165 trials of the Fib grammar whereas they only had 600 trials in the present experiment. Third, Schmid et al. (2023a) involved 159 participants whereas ours had only 53, which may be responsible for weaker statistical power, knowing that indeed the effect of ambiguity at level 3 was weaker than that at other levels in Schmid et al. (2023a).

3.3 Experiment 2 : Switching from Skip to Fib

In Experiment 2, participants were exposed to 5 blocks of the Skip grammar followed by one block of the Fib grammar. The first aim of Experiment 2 was to test prediction (3) of the k-points hypothesis according to which identification of k-points in the final Fib block should be facilitated compared to the last block of Skip because participants would have already been exposed to the surface regularity "*a 0 is always followed by a 1*" that permits identification of k-points. If this prediction is correct, k-points should show faster RTs and better accuracy in the Fib block compared to the last Skip block. The second aim of this experiment was to test prediction (b) of the recursive merge hypothesis that participants extract a hierarchical structure in the processing of the Skip grammar in the same way as in Fib, since both exhibit an underlying hierarchical structure.

3.3.1 Methods

3.3.1.1. Participants

Fifty-one students (19 men and 32 women; mean age 21.8 years old) from an introductory psycholinguistics course from the university of Geneva participated in the experiment for course credits. All participants reported normal or corrected-to-normal vision. This study was approved by the ethics commission of the University of Geneva. Informed consent was obtained from each of the participants included in the study.

3.3.1.2 Materials

The training sequence was identical to that of Experiment 1. The sequences used in the experimental blocks corresponding to the Skip grammar were created by extracting 255 sub-sequences of 120 points from generation 12 of the Skip grammar. Each of those 255 sub-sequences corresponded to a unique block and were used only once across participants. The sequences used in the experimental blocks corresponding to the Fib grammar were created by extracting 51 sub-sequences of 120 points from generation 21 of the Skip grammar. Each of those 51 sub-sequences corresponded to a unique block and were used only once across participants.

3.3.1.3 Design and procedure

The design and procedure were identical to Experiment 1 except that the order of the trials in experimental blocks 1 to 5 was determined by the Skip grammar and in experimental block 6, the order of the trials was determined by the Fib grammar. Stimuli were presented electronically using the E-Prime 2.0 software *(Psychology Software Tools, Inc. (2016)*.

3.3.1.4 Data analyses

Data cleaning was identical to Experiment 1. We removed one participant who had an error rate superior to 3 SD to the mean error rate in two blocks. We also removed 30 trials where participants did not respond after 1200 ms. Analyses for the remaining 50 participants were identical to Experiment 1.

3.3.2 Results

3.3.2.1 Processing of k-points

Analyses of RTs showed no effect of *Grammar* (β = 9.17, *SE* = 5.07, t = 1.809, p = .070). The effect of *Grammar* on accuracy was also not significant (β = 0.22, *SE* = 0.12, z = 1.876, p = .061).

3.3.2.2 Processing of Skip hierarchical structure

Processing of surface statistical regularities (Level 0).

Analyses of RTs showed a main effect of *Exposure* (β = -17.35, *SE* = 0.56, t = -30.60, p < .001) with a mean reduction of 87 ms from block 1 to block 5. There was also a main effect of *Skip ambiguity* $level_0$ (β = -62.17, *SE* = 1.88, t = -33.15, p < .001) with disambiguated points being faster than non-disambiguated ones by 62 ms. The interaction *Exposure*Skip ambiguity* $level_0$ was also significant (β = -10.97, *SE* = 1.32, t = -8.296, p < .001) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -80 ms) than non-disambiguated points ($M_{block5 - block1}$ = -32 ms). Results are shown in Fig 5.

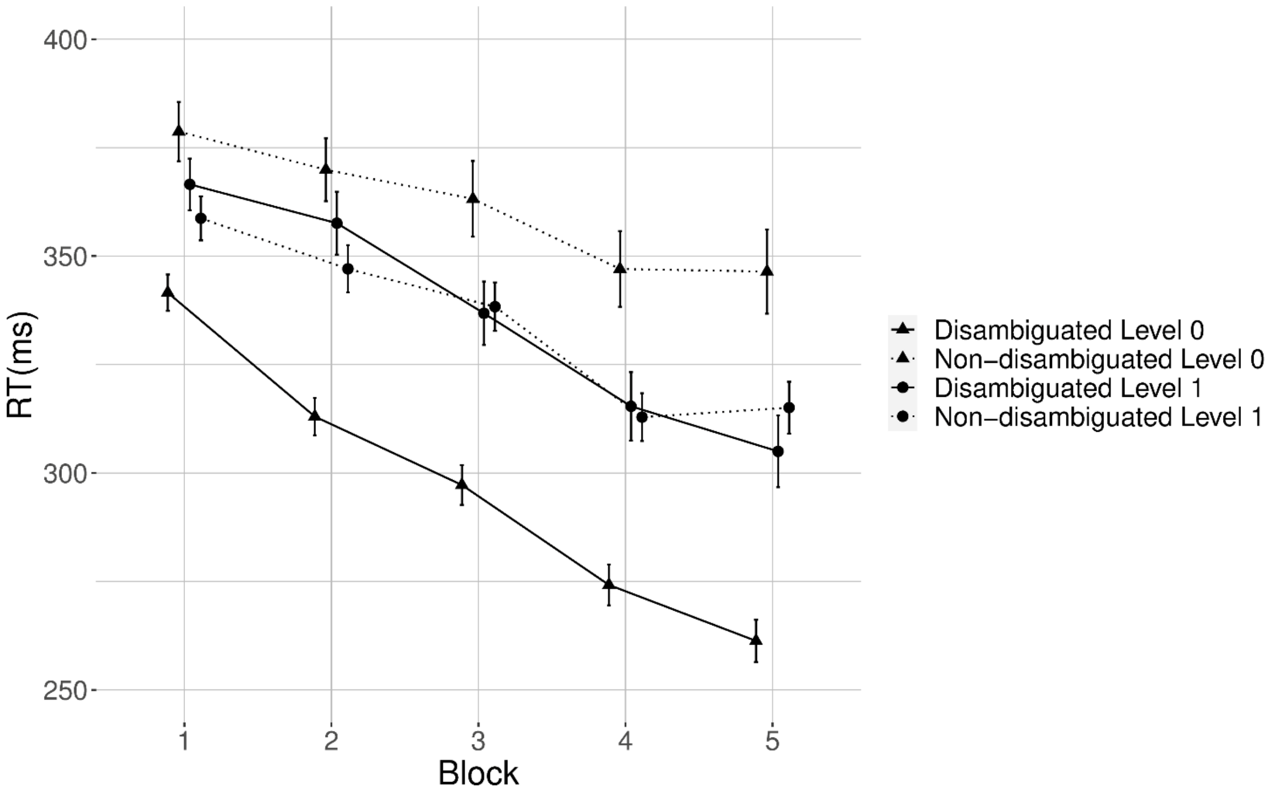


Fig 5. Experiment 2: mean RTs for disambiguated and non-disambiguated points by block (level 0 and 1). Errors bars denote the 95% confidence interval.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.10, *SE* = 0.03, *z* = -3.483, *p*< .001) with a mean reduction of accuracy of 1.6 % from block 1 to block 5. There was also a main effect of *Skip ambiguity level*$_0$ (β = 2.74, *SE* = 0.10, *z* = 27.80, *p*< .001) with higher accuracy for disambiguated points (*M* = 0.99) than for non-disambiguated points (*M* = 0.87). The effect of *Exposure* significantly interacted with *Skip ambiguity level*$_0$ (β = 0.42, *SE* = 0.07, *z* = 5.898, *p* < .001) with accuracy increasing for disambiguated points over time ($M_{block5 - block1}$ = 0.01) and decreasing for non-disambiguated points ($M_{block5 - block1}$ = -0.09). Results are shown in Table 2.

Table 2

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Disambiguated and Non-Disambiguated Points in the Skip grammar by Hierarchical Levels and Blocks.

		Block 1		Block 2		Block 3		Block 4		Block 5	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 0	Disambiguated	0.98	0.13	0.99	0.11	0.99	0.08	0.99	0.10	0.99	0.07
	Non-disambiguated	0.91	0.28	0.88	0.33	0.86	0.35	0.86	0.35	0.82	0.38
Level 1	Disambiguated	0.96	0.22	0.97	0.18	0.97	0.18	0.97	0.17	0.97	0.18
	Non-disambiguated	0.97	0.17	0.96	0.19	0.96	0.20	0.96	0.20	0.95	0.22
Level 2	Disambiguated	0.97	0.18	0.97	0.17	0.97	0.18	0.96	0.19	0.95	0.22
	Non-disambiguated	0.98	0.15	0.95	0.21	0.95	0.22	0.95	0.22	0.95	0.22
Level 3	Disambiguated	0.92	0.27	0.90	0.31	0.88	0.32	0.89	0.31	0.86	0.35
	Non-disambiguated	0.88	0.32	0.83	0.38	0.78	0.41	0.76	0.43	0.74	0.44
Level 4	Disambiguated	0.98	0.13	0.95	0.22	0.94	0.25	0.94	0.23	0.93	0.26
	Non-disambiguated	0.97	0.16	0.96	0.20	0.96	0.21	0.95	0.21	0.96	0.20

Hierarchical processing at level 1.

Analyses of RTs showed a main effect of *Exposure* (β = -13.58, *SE* = 0.63, t = -21.428, p < .001) with a mean reduction of 53 ms from block 1 to block 5. There was no effect of S*kip ambiguity level*$_1$ (β = 2.37, *SE* = 1.86, t = 1.275, p = .203), but a significant interaction *Exposure*Skip ambiguity level*$_1$ (β = -4.35, *SE* = 1.31, t = -3.31, p < .001) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -62 ms) than non-disambiguated points ($M_{block5 - block1}$ = -44 ms). Results are shown in Fig 5. Concerning accuracy, we found a significant effect of *Exposure* (β = -0.07, *SE* = 0.03, z = -2.117, p = .034) with a mean reduction of accuracy of 1.1 % from block 1 to block 5. There was a main effect of *Skip ambiguity level*$_1$ (β = 0.22, *SE* = 0.10, z = 2.146, p = .032) with accuracy higher for disambiguated points (M = 0.97) than for non-disambiguated points (M = 0.96). The interaction *Exposure*Skip ambiguity level*$_1$ was also significant (β = 0.16, *SE* = 0.07, z = 2.300, p = .021) with accuracy increasing for disambiguated points over time ($M_{block5 - block1}$ = 0.005) and decreasing for non-disambiguated points ($M_{block5 - block1}$ = -0.02). Results are shown in Table 2.

Hierarchical processing at level 2.

Analyses of RTs showed a main effect of *Exposure* (β = -11.96, *SE* = 0.80, *t* = -14.872, *p* < .001) with a mean reduction of 60 ms from block 1 to block 5. There was no significant effect of *Skip ambiguity level₂* (β = -3.26, *SE* = 2.31, *t* = -1.415, *p* = .157), but a significant interaction *Exposure*Skip ambiguity level₂* (β = -5.42, *SE* = 1.63, *t* = -3.330, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -52 ms) than non-disambiguated points ($M_{block5 - block1}$ = -32 ms). Results are shown in Fig 6. Concerning accuracy, we found a main effect of *Exposure* (β = -0.13, *SE* = 0.04, *z* = -3.234, *p* = .001) with a mean reduction of accuracy of 2.2 % from block 1 to block 5. There was no effect of *Skip ambiguity level₂* (β = 0.20, *SE* = 0.12, *z* = 1.745, *p* = .081). The interaction *Exposure*Skip ambiguity level₂* was also not significant (β = 0.03, *SE* = 0.08, *z* = 0.393, *p* = .695). Results are shown in Table 2.

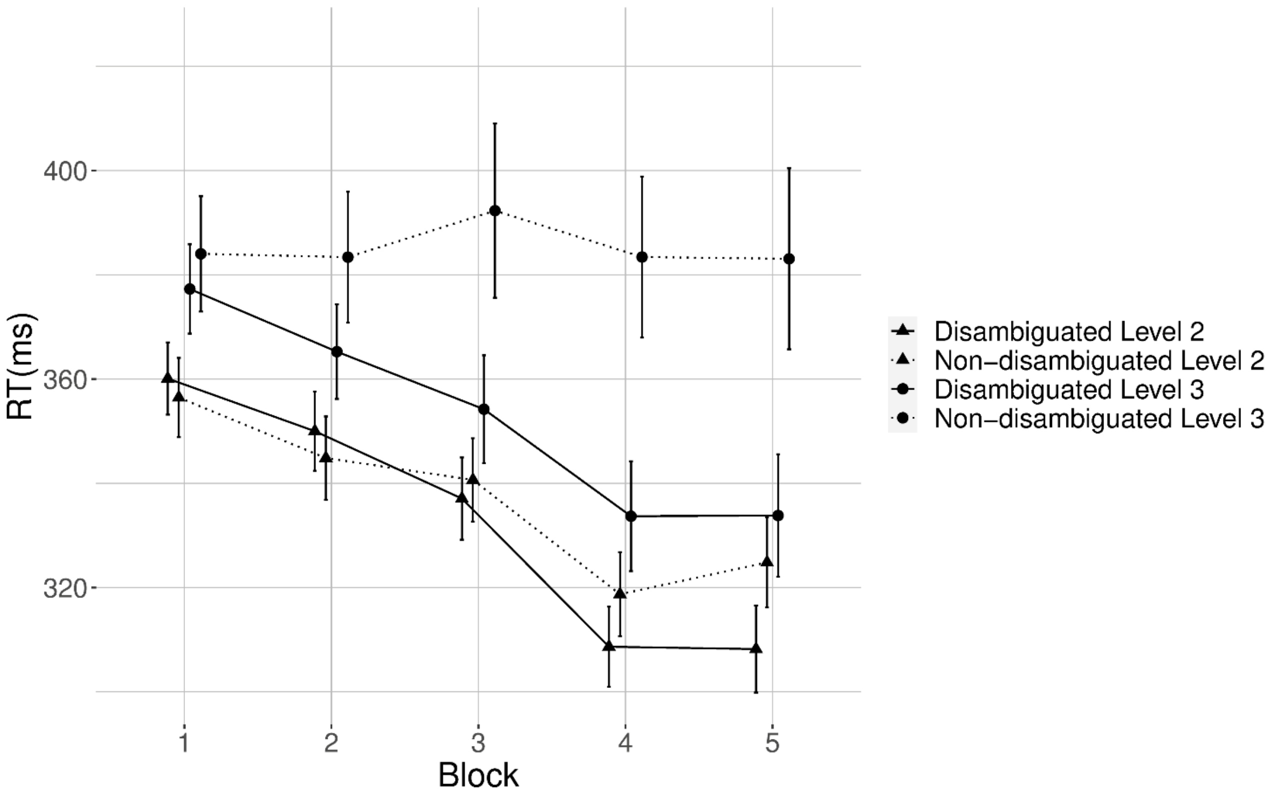


Fig 6. Experiment 2: mean RTs for disambiguated and non-disambiguated points by block (level 2 and 3). Errors bars denote the 95% confidence interval.

Hierarchical processing at level 3.

Analyses of RTs showed a main effect of *Exposure* (β = -8.98, *SE* = 1.19, t = -7.542, p < .001) with a mean reduction of 45 ms from block 1 to block 5. There was a main effect of S*kip ambiguity level*$_3$ (β = -25.397, *SE* = 3.93, t = -6.466, p < .001) with disambiguated points being faster than non-disambiguated ones by 24 ms. The interaction *Exposure*Skip ambiguity level*$_3$ was also significant (β = -11.08, *SE* = 2.76, t = -4.01, p < .001) with a more important reduction over time for disambiguated points ($M_{block5 - block1}$ = -43 ms) than non-disambiguated points ($M_{block5 - block1}$ = -1 ms). Results are shown in Fig 6. Concerning accuracy, we found a main effect of *Exposure* (β = -0.18, *SE* = 0.03, z = -5.426, p < .001) with a mean reduction of accuracy of 9 % from block 1 to block 5. There was a main effect of *Skip ambiguity level*$_3$ (β = 0.75, *SE* = 0.09, z = 7.950, p < .001) with higher accuracy for disambiguated points (M = 0.89) than for non-disambiguated points (M = 0.80). The interaction *Exposure*Skip ambiguity level*$_3$ was not significant (β = 0.09, *SE* = 0.07, z = 1.268, p = .205). Results are shown in Table 2.

Hierarchical processing at level 4.

Analyses of RTs showed a main effect of *Exposure* (β = -9.19, *SE* = 1.20, t = -7.642, p < .001) with a mean reduction of 46 ms from block 1 to block 5. There was no effect of *Skip ambiguity level*$_4$ (β = 5.83, *SE* = 3.56, t = 1.636, p = .102), but a significant interaction *Exposure*Skip ambiguity level*$_4$ (β = 5.01, *SE* = 2.51, t = 1.997, p = .046) with a less important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -22 ms) than non-disambiguated points ($M_{block5 - block1}$ = -38 ms). Results are shown in Fig 7. Concerning accuracy, we found a main effect of *Exposure* (β = -0.16, *SE* = 0.06, z = -2.464, p = .014) with a mean reduction of accuracy of 2.8 % from block 1 to block 5. There was no effect of *Skip ambiguity level*$_4$ (β = -0.30, *SE* = 0.18, z = -1.694, p = .090). The interaction *Exposure*Skip ambiguity level*$_4$ was also not significant (β = -0.16, *SE* = 0.13, z = -1.286, p = .199). Results are shown in Table 2.

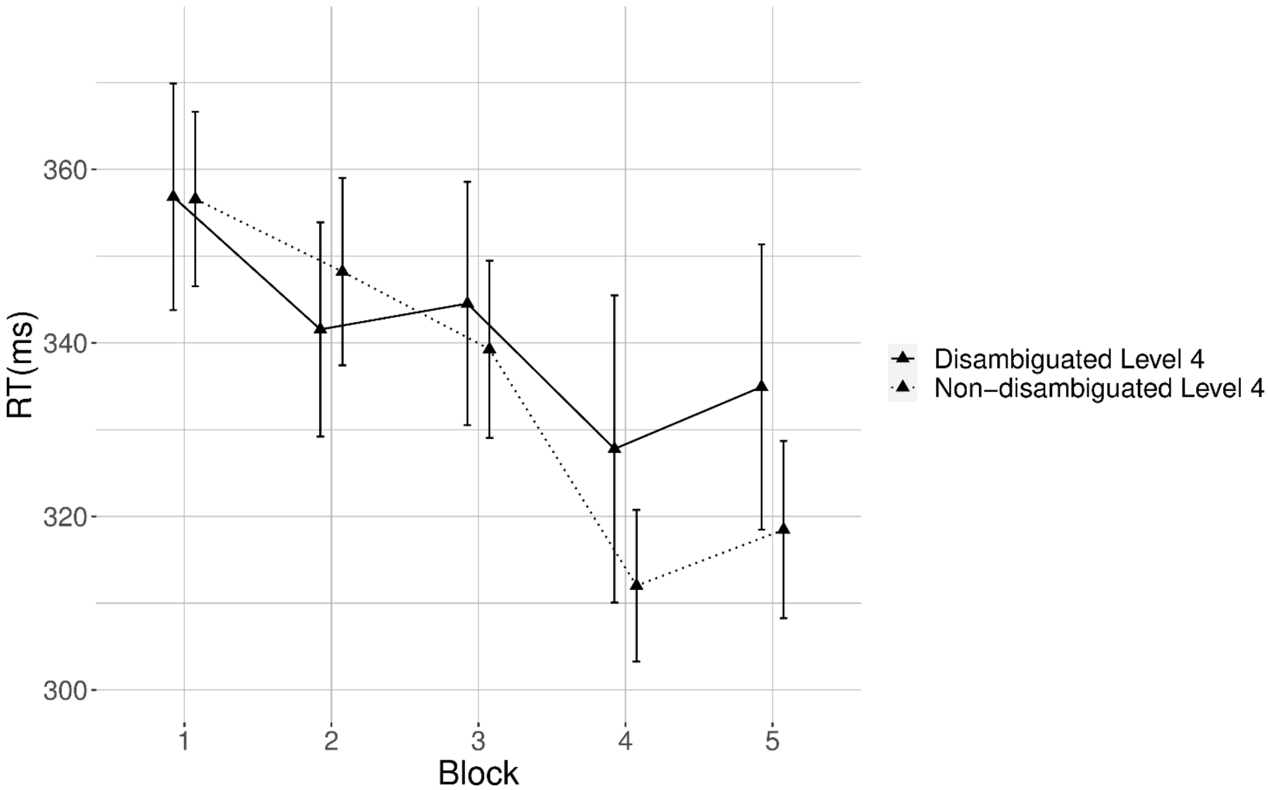


Fig 7. Experiment 2: mean RTs for disambiguated and non-disambiguated points by block (level 4). Errors bars denote the 95% confidence interval.

3.3.2.3 Comparison of Experiment 1 and 2 : processing of k-points in the Fib grammar and the Skip Grammar

In order to evaluate prediction (4) of the k-points hypothesis according to which the processing of k-points in the Fib grammar should be facilitated compared to k-points in the Skip grammar independently of prior exposure to one of the grammars, we compared the RTs and accuracy of k-points in the Fib blocks of Experiment 1 to that of the Skip blocks of Experiment 2. The model included two fixed-effect factors and their interaction: *Exposure, Grammar,* and *Exposure*Grammar*. *Exposure* was modeled in the same way as in the previous analysis. *Grammar* is a discrete variable contrasting the k-points of the Fib blocks of Experiment 1 to that of the Skip blocks of Experiment 2. As random effects, the models had intercepts for *Participants.* The modality “Fib” of the factor *Grammar* was set as the intercept of the model.

Analyses of RTs showed a main effect of *Exposure* (β = -6.76, *SE* = 0.67, *t* = -10.158, *p* < .001) with a mean reduction of 28 ms from block 1 to block 5. There was no effect of *Grammar* (β = -1.57, *SE*

= 10.81, t = -0.146, p = .885). The interaction *Exposure*Grammar* was however significant (β = -3.15, SE = 1.38, t = -2.283, p = .022) with a more important reduction over exposure for Skip k-points ($M_{block5\,-\,block1}$ = -32 ms) than Fib k-points ($M_{block5\,-\,block1}$ = -23 ms). Concerning accuracy, we found a main effect of *Exposure* (β = -0.18, SE = 0.02, z = -8.308, p < .001) with a mean reduction of accuracy of 7 % from block 1 to block 5. There was a main effect of *Grammar* (β = -0.56, SE = 0.14, z = -3.983, p < .001) with accuracy higher for k-points of the Fib grammar (M = 0.92) than for k-points of the Skip grammar (M = 0.87). The interaction *Exposure*Grammar* was however not significant (β = -0.001, SE = 0.04, z = -0.044, p = .965).

3.3.3 Discussion

The first aim of Experiment 2 was to evaluate prediction (3) of the k-points hypothesis that prior exposition to the surface regularity "*a 0 is always followed by a 1*" in the Skip grammar should facilitate the ulterior identification of k-points in the Fib grammar. There was no difference in RTs or accuracy between Fib k-points and Skip k-points, contrary to Experiment 1 where Skip k-points were processed more slowly and led to more errors than Fib k-points. These results suggest that switching from Skip to Fib, unlike the switch from Fib to Skip, has no detrimental effect on k-points processing. However, capitalizing on a null result is always tricky and more data is needed to interpret this effect further. In any case, contrary to prediction (3) of the k-points hypothesis, prior processing of the Skip grammar did not facilitate identification of k-points in Fib.

The second aim of Experiment 2 was to evaluate prediction (b) that participants extracted a hierarchical structure during the processing of the Skip grammar, as they did for the Fib grammar, through the recursive combination of elements spanning across deterministic transitions. The results of the Skip blocks showed that for levels 0, 1, 2 and 3, RTs decreased significantly more for the disambiguated points than for their non-disambiguated counterparts. The accuracy was also systematically higher for disambiguated points than for non-disambiguated points, suggesting that the decrease in RTs was not due to a speed accuracy trade-off. The results of levels 0, 1, 2 and 3 are therefore in line with the prediction of the recursive merge account.

According to prediction (4) of the k-points hypothesis, a general processing advantage was expected for k-points in Fib compared to k-points in Skip. Comparison between Fib processing (Experiment 1) and Skip processing (Experiment 2) showed that RTs of k-points in Skip decreased more through exposure than k-points in Fib. Although Fib k-points were found to be more accurately processed than Skip k-points, this effect was already present in the first block and did not change with exposure, suggesting that the difference between the two grammars did not arise as a consequence of learning. These results therefore invalidate prediction (4) of the k-points hypothesis: the lack of isomorphism did not hinder the learning of k-points in the Skip grammar.

Finally, an intriguing finding was found at level 4 at which RTs decreased more strongly for non-disambiguated points than for disambiguated points. There was no difference in accuracy at this level. We discuss the implication of this incongruent result in the General discussion.

3.4 General Discussion

The aim of the present study was to evaluate two proposals that account for the extraction of hierarchical structure in the processing of the Fib grammar. The k-points hypothesis (Krivochen et al., 2018; Vender et al., 2020) proposes that within the Fib grammar, the identification of certain points, called k-points, would allow the construction of the local hierarchical structure of the sequence due to their specific structural status. In order to identify these k-points, the parser would make the assumption that the linear order "*a 0 is always followed by a 1*" is symmetric to the derivational order "*a k-point is dominated by a 0*". This assumption would be possible in the Fib grammar because of the isomorphism between the surface properties of the sequence and its derivational properties. In support of this hypothesis, Vender et al. (2020) reported a processing advantage for k-points in the Fib grammar compared to k-points in an alternative grammar, Skip, which has the same surface statistical regularities as Fib but without isomorphism between the linear and derivational orders. An alternative hypothesis attributing no role to isomorphism has been proposed to explain hierarchical learning in the Fib grammar. The recursive merge hypothesis Schmid et al. (2023a) proposes that the process behind the extraction of hierarchical structure consists in the progressive recursive combination of points that are linked by deterministic transitions resulting in a complex multi-level hierarchical structure composed of embedded constituents. Although the recursive merge hypothesis relies on recursion, another formal property of Fib, and not on isomorphism, it does not rule out the possibility that isomorphism also plays a role in structure extraction.

Since the key observation in favor of the k-points hypothesis was based on the comparison between Fib and Skip, and the predictions of the recursive merge hypothesis in Skip had not been explored, we conducted the present study to more systematically explore the predictions of the two hypotheses for these two grammars. Two experiments were reported using an SRT task. In Experiment 1, participants performed five blocks of the Fib grammar followed by one block of the Skip grammar, while the opposite order was tested in Experiment 2. We review the results in light of the predictions of each hypothesis, and then discuss a possible explanation for some of the unexpected results, lying in the differing type of self-similarity of our two test grammars.

3.4.1 Predictions of the two hypotheses in light of the data

3.4.1.1 Predictions of the k-points hypothesis

Prediction (1)

A k-point in the Fib grammar is defined as a 1 that is dominated by a 0 at generation n-1. Thus, all k-points have the same formal status in the Fib grammar. If identification of k-points is the sole factor that drives hierarchical structure building, then all k-points should be processed in the same way in the Fib grammar. We found that at level 2, RTs decreased more through exposure for disambiguated than non-disambiguated points whereas accuracy decreased more for non-disambiguated points than for disambiguated points. Since at this level both disambiguated and non-disambiguated points are k-points, the data do not support prediction (1).

Prediction (2)

The k-points hypothesis predicted that after exposure to the Fib grammar, the processing of k-points should be impaired in the Skip grammar because the Skip grammar does not show the isomorphism found in Fib. Our results are in line with this prediction : the k-points in the Skip block following Fib blocks in Experiment 1 resulted in longer RTs and a higher error rate than the k-points in the last Fib block.

Prediction (3)

The k-points hypothesis predicted that prior exposition to the surface regularity "*a 0 is always followed by a 1*" in the Skip grammar should facilitate the ulterior identification of k-points in the Fib grammar. We found no difference in processing between Skip k-points and Fib k-points either in RTs or in accuracy. Thus, exposure to the Skip grammar did not facilitate further processing of k-points in the Fib grammar. However, our results showed an asymmetry between the transition from one grammar to the other. In Experiment 1, the switch from Fib to Skip led to a detrimental effect on k-points processing, while it was not the case in the switch from Skip to Fib. Further studies are needed to investigate this question. In any case, prediction (3) of the k-points hypothesis is invalidated.

Prediction (4)

The k-points hypothesis stated that k-points should be easier to process in the Fib grammar compared to the Skip grammar, independently of prior exposition. Results showed that RTs decreased more through exposure for k-points in the Skip grammar compared to k-points in the Fib grammar, invalidating prediction (4). The lack of isomorphism in the Skip grammar does impair the processing of k-points.

3.4.1.2. Predictions of the recursive merge hypothesis

Prediction (a)

The recursive merge hypothesis predicted that the processing of the Fib grammar should result in the progressive anticipation of the points that follow a higher order deterministic transition. The results of Experiment 1 are consistent with this hypothesis: across exposure, disambiguated points showed a decrease in RTs and an increase in accuracy greater than non-disambiguated points at levels 0, 1 and 2. Those results replicate Schmid et al. (2023a). However, contrary to Schmid et al., (2023a), we did not find an effect at level 3. As previously discussed, this null result may be due to the lower statistical power and/or the longer RSI of our experiment.

Prediction (b)

The recursive merge hypothesis assumes that the same mechanism allows for the extraction of the hierarchical structure in Skip and Fib. Results from Experiment 2 on Skip sequences showed that RTs and error rates of the disambiguated points of hierarchical levels 0, 1, 2 and 3 decreased more strongly than those of their non-disambiguated counterparts, which parallels what was found for Fib sequences in Experiment 1. In line with the recursive merge hypothesis, those results suggest that participants built a hierarchical structure up to the third level. They also show that the isomorphism between the linear and derivational order is not a necessary condition for the extraction of a hierarchical structure since the Skip grammar lacks this isomorphism. Nevertheless, the recursive merge hypothesis fails to account for the finding that RTs decreased more through exposure for the non-disambiguated points than for the disambiguated points at level 4. We discuss a potential ad-hoc explanation of this finding in section 3.4.2.

In summary, with the exception of prediction (2) which replicates the results of Vender et al. (2020), the k-points hypothesis seems unable to account for the findings of the present study. However, the finding that disambiguated points were globally anticipated better than their non-disambiguated counterparts, in both the Fib and Skip grammars, supports the hypothesis that a common hierarchical building mechanism was used that consists in the recursive merge of deterministic transitions. Our results therefore suggest that the absence of isomorphism does not prevent participants from elaborating a hierarchical structure in the Skip grammar. Moreover, the similarities between the processing of each grammar cast doubt on the hypothesis that the isomorphism in Fib is the factor driving hierarchical elaboration. Nevertheless, the reverse ambiguity effect found at Skip level 4 suggests processing differences between the two grammars. What are the differences between the two grammars that could explain this effect ?

In the next section, we reinterpret the k-points hypothesis in suggesting that the presence or absence of isomorphism actually reflects a difference in the form of self-similarity between Fib and Skip. We then propose that this variation in self-similarity affects the recursive merge building process through the interplay between self-similarity and signal complexity.

3.4.2 The role of self-similarity

The isomorphism between surface properties (i.e., horizontal order) and derivational properties (i.e., vertical order) is at the heart of the k-points hypothesis. It is expressed in the Fib k-points: their formal status recapitulates the structure of the grammar. This reflects the fact that Fib shows *perfect structure preservation*: the structure of this grammar is entirely scale-free (Krivochen et al., 2018). Skip does not exhibit the isomorphism of Fib; however, both grammars have some common properties. This is not surprising since the rewriting rules of Skip correspond to Fib natural constituents. Thus, Skip inherits some properties of Fib, namely recursive constituent structure, asymmetry and aperiodicity. However, the Skip grammar departs from Fib in an important way : in Skip, transitional probabilities do not always match those of higher order constituents (Fig 2. C). This is because Skip does not show perfect structure preservation : the self-similarity of Skip is not uniform across the hierarchy but is *bounded*. Thus, in contrast to Fib, Skip's self-similarity is not scale-free but operates in a restricted range. We will refer to *weak* self-similarity (bound, as in Skip) and *strong* self-similarity (scale-free, as in Fib).

We hypothesize that self-similarity acts as reinforcing the representation of the hierarchical structure. In Fib, due to its strong self-similarity, a disambiguated point at level n always predicts a deterministic transition at level n-1; in other words, the creation of a new level can always be done by predicting a point/constituent that was itself used to predict a point/constituent at the level immediately below. As a result, in Fib, ambiguous points are always integrated in the structure in the same way, and the representation of the structure thus gets reinforced with time. In Skip, a disambiguated point at level n does not always predict a deterministic transition at level n-1. In consequence, the robustness of the hierarchical structure is weaker in Skip.

Interestingly, the type of self-similarity covaries with another parameter that may play a key role in sequence processing: complexity. Indeed, as we saw in the introduction, the level of complexity of the input leads the cognitive system to compress it into a more abstract format (Planton et al., 2021; Pothos, 2010; Radulescu et al., 2019, 2021). When the complexity of the input exceeds the processing capacity of the parser, it compresses the signal to reduce processing cost. In that regard, hierarchical elaboration can be seen as a way to reduce the entropic state of the parser. Skip and Fib differ in their complexity: at the surface level, Skip is less complex than Fib because the transitional probabilities are more asymmetric (i.e., the closer the different transitional probabilities are, the more random, and therefore the more complex is the signal). However, due to its strong self-similarity, the elaboration of a new hierarchical level in Fib decreases the entropic state of the parser at a constant rate, i.e., the proportion of ambiguous points that are disambiguated is identical at each level. Conversely, this rate is not equivalent within each level in Skip because of its weak self-similarity. This means that the complexity of Skip varies across the hierarchy: some levels are less complex than others. It might thus be that when the parser reaches some levels, its entropic state is reduced enough so that the input does not trigger further compression, thus halting the process of hierarchical elaboration. Alternative parsing strategies that were too costly to be considered at the beginning of the learning phase might then become available. Those strategies would operate on the hierarchical structure previously built but they would not consist of further compression of the representation of the signal.

This interplay between complexity and self-similarity may explain the unexplained learning advantage for the non-disambiguated points of level 4 in Skip. At level 3, the parser would have resorted to a strategy that is a priori more costly in terms of cognitive load but made accessible by the fact that the entropic state of the parser has been sufficiently reduced at that level. Let's consider the distribution of constituents at level 3 of the Skip grammar (Fig 2C). At this level, [01] is always followed by [10101], [10101] is followed by [01] in 73% of the cases, and [10101] is followed by [10101] in 27% of the cases. Therefore, the repetition of the constituent [10101] occurs rarely (about 5 times per block). This means that except for the 5 times where [10101] is repeated, the rest of the sequence consists of a periodic alternation between [01] and [10101]. Thus, when level 3 is reached, the parser could hypothesize that the sequence consists of the periodic alternation of these two constituents. This would be a relatively efficient strategy because it is only violated five times per block, i.e., when constituent [10101] is repeated. According to this hypothesis, the repetition of [10101] should give rise to a slowdown. This is what our results show: the non-disambiguated points of level 3, which correspond to the first point of [10101] when it is repeated (i.e., the 1 in bold in [10101][**1**0101]), do not give rise to any learning ($M_{block5 - block1}$ = -1 ms). In order to explain the incongruent effect at level 4, let’s examine where the disambiguated and non-disambiguated points of this level occur. Below, we have put in bold the disambiguated points (i) and the non-disambiguated points (ii) of level 4 :

(i) [10101] [10101] [**0**1]

(ii) [01] [10101] [**0**1]

We can see in (i) that disambiguated points of level 4 systematically follow [10101] when it repeats, whereas (ii) show that non-disambiguated points of level 4 systematically follow [10101] when it does not repeat. Therefore, if the periodic alternation structure developed by the parser is less robust because of weak self-similarity, the unexpected repetition of the constituent [10101] may trigger the revision of that structure. In consequence, the parser would be cautious about the points that follow the repetition of [10101], thus explaining the slower RTs for the disambiguated points of level 4.

In a nutshell, the proposed hypothesis relies on the core underlying assumption of the k-points hypothesis, which is that the type of self-similarity between the grammars diverge. This property is assumed to interact with the complexity of the parser's signal representation. Skip not being uniformly self-similar, the parser could reach a sufficient degree of signal compression that makes available other strategies that do not require deepening the hierarchical structure. In contrast, Fib being scale-free, the error from the signal always reinforces the structure and triggers further hierarchical elaboration. The apparent opposition between the hypotheses of Schmid et al. (2023a) and Vender et al. (2020) would thus be due to the fact that the measure on which Vender et al. (2020) relied (i.e., the processing of k-points) is too restricted to capture the impact of the formal properties of Fib and Skip on the mechanisms put in place by the participants.

We have proposed that the self-similarity of the signal plays a role along with signal complexity. One possibility could be that self-similarity has the effect of reinforcing the existing structure while complexity leads to further compression of that structure. In this view, the two parameters are complementary. Our study does not allow us to conclude on this question because as explained above, the Skip grammar and the Fib grammar vary both in complexity and in the form of their self-similarity. Further research where self-similarity and input complexity are carefully teased apart are thus necessary.

Chapter 4. Non-linear effects of presentation rate on sequence learning

4.1 Introduction

The understanding of the mechanism allowing the acquisition of the regularity underlying a sequence of events represents a fundamental question in cognitive science. Among the multiple methods adopted to explore sequence learning, the Serial Reaction Time (SRT) task is one of the most prominent. In this task, the elements of the sequence are presented one by one and the sequence determines the position of a stimulus on a screen. On each trial, the participant must press the button associated with the position of the stimulus as quickly as possible. Once the response is made, the stimulus disappears and the next trial begins. In the initial version of Nissen and Bullemer (1987), the stimulus could appear in four possible positions; this setting has been widely adopted in following studies. The assessment of sequence learning is done in two phases. In the learning phase, participants perform several blocks where the order of appearance of the stimuli is determined by the target sequence to be learned. This phase is followed by a so-called *transfer block* where the order of the stimuli is determined by an alternative sequence and then again by a block following the target sequence. The target sequence is considered as learned if the participants show a *transfer effect*, i.e., a slowing down in the transfer block compared to the adjacent sequenced blocks (Schwarb & Schumacher, 2012). The target sequence is typically 10-12 items long and is presented in a loop through the blocks. The most commonly used type of sequence follows a Second Order Conditional (SOC) structure (Reed & Johnson, 1994). In SOC sequences, the position of a trial t_n can be predicted with certainty by the position of the two preceding trials ($p(t_n|t_{n-1},t_{n-2}) = 1$) although the position of trial t_{n-1} alone is non-informative ($p(t_n|t_{n-1}) = .25$). In most studies published after Reed and Johnson (1994), the alternative sequence used in the transfer block also follows the SOC structure (but with a different surface expression) instead of the random sequence used previously. This allows for precise control of the statistical distribution of the alternative sequence and thus ensures that if a transfer effect is observed, it is due to the fact that the participants have learned the target sequence (see Reed & Johnson, 1994, for a justification of this manipulation).

Many parameters influencing sequence learning in the SRT task have been reported, such as sequence structure, alignment between stimulus locations and response key, or the presence of a secondary task (see Forkstam & Petersson, 2005; Schwarb & Schumacher, 2012; for reviews). However, the influence of yet another parameter, the duration of the Response-to-Stimulus Interval (RSI), remains unclear. Three hypotheses on the influence of RSI duration on sequence learning have been proposed in the literature. The first hypothesis states that the RSI would affect the amount of knowledge acquired about the sequence because of its impact on information processing in Working Memory (WM) (Frensch & Miner, 1994; Soetens et al., 2004). RSI duration would affect learning through the decay of stimulus representations in working memory : the shorter the duration of the RSI, the better the sequence would be learned. For the sake of clarity, we will refer to this hypothesis as the "Decay hypothesis". The second hypothesis states that the modulation of the RSI would not affect the learning of the sequence *per se*, but rather the performance in the SRT task (Willingham et al., 1997). According to this hypothesis, the duration of the RSI would affect the preparation of the response. Learning would be relatively equivalent across different RSI durations but could only be detected when the RSI is sufficiently short. With a long RSI, when the stimulus appears at an unexpected position in the transfer block, participants would have enough time to inhibit the learned response which would hide the transfer effect. We will refer to this hypothesis as the "Preparation hypothesis". Finally, the third hypothesis proposes that the RSI would mainly influence the ability to elaborate a conscious representation of the acquired knowledge (Cleeremans & Sarrazin, 2007; Destrebecqz & Cleeremans, 2001, 2003; Frensch & Miner, 1994; Huang et al., 2017; Kuhn & Dienes, 2006; Norman et al., 2007; Savalia et al., 2016; Soetens et al., 2004; Verwey & Dronkers, 2019; Verwey & Wright, 2014; Willingham et al., 1997).

The idea being that participants could only acquire explicit knowledge of the sequence when the duration of the RSI is long enough. When the RSI is very short or null, the knowledge acquired would remain largely implicit. Since the acquisition of item-based representations is limited to implicit learning, and acquiring more abstract knowledge requires explicit representation of the sequence, the duration of the RSI would affect the degree of abstraction of the acquired knowledge. We will refer to this hypothesis as the "Awareness hypothesis".

In what follows, we first review the empirical evidence for each of these hypotheses about the role of the RSI in the SRT task. To date, no consensus has been reached and all the aforementioned hypotheses remain plausible candidates. We then argue that the limitations of the method classically used to assess sequence learning (i.e., the transfer effect) make it unsuitable to settle between these hypotheses. We therefore assessed learning without using a transfer block. To achieve this, we took advantage from the specific properties of the sequence generated by the Fibonacci grammar which enable to measure learning continuously without having to compare performance to a transfer block.

Hypothesis 1: RSI duration affects sequence learning

The Decay hypothesis is based on the observation in some studies that the magnitude of the transfer effect decreases with increasing RSI duration. Frensch and Miner (1994) observed that the magnitude of the transfer effect was smaller for an RSI of 1500 ms than for an RSI of 500 ms. This detrimental effect of RSI lengthening on the transfer effect was replicated by other studies (Soetens et al., 2004; Stadler, 1995; Willingham et al., 1997). Frensch and Miner (1994) proposed that this detrimental effect would be due to the decay of the stimulus representation in WM. With a long RSI, the number of stimuli simultaneously active in WM would decrease, which would reduce the detection of sequence regularity and thus decrease the magnitude of the transfer effect. Nevertheless, the influence of the RSI on the magnitude of the transfer effect has not always been replicated, with multiple studies reporting no effect even with very different RSI values (Destrebecqz & Cleeremans, 2001, 2003; Huang et al., 2017; Norman et al., 2007) and none showing the opposite effect. For example, Destrebecqz and Cleeremans (2003) found that the magnitude of the transfer effect was identical for RSIs of 0 ms, 250 ms and 1500 ms. Moreover, although the claim that memory traces decay over time is assumed by multiple models of working memory (Barrouillet et al., 2004; Hommel, 1994; S. T. Mueller et al., 2003), it is not consensual due to the fact that effects typically attributed to decay can also actually be due to interference (Lewandowsky & Oberauer, 2009; Oberauer, 2013, 2019; Oberauer & Lewandowsky, 2013, 2014; Ricker et al., 2016).

Hypothesis 2: RSI duration would not affect sequence learning but response preparation

The Preparation hypothesis provides an alternative explanation for the finding of weaker transfer effects with long RSIs in the SRT task. Willingham et al. (1997) manipulated the duration of the RSI within participants. In the long-short condition, participants first performed four training blocks and a transfer block with a 1500 ms RSI. After this, participants performed another training block with the same target sequence but with an RSI reduced to 500 ms, followed by a final transfer block. In the short-long condition, the short RSI was initially presented, followed by the long RSI. Participants in the long-short condition showed no transfer effect when the RSI lasted 1500 ms. In contrast, the same participants displayed a transfer effect once the RSI duration was reduced to 500 ms, even though they performed only one training block at this presentation rate. To explain these results, Willingham et al (1997) suggested that sequence learning occurred when the RSI was 1500 ms long, but that the long presentation rate hid the learning effect, which was only visible when the RSI was short enough. The decrease in the magnitude of the transfer effect with increasing RSI duration frequently found in the literature would thus not reflect sequence learning but better response preparation. With a longer RSI, participants would have more time to prepare for the next trial, so the surprise effect induced by a stimulus arriving at an unexpected position would be less, resulting in a reduced transfer effect.

Another observation suggesting that RSI duration does not affect sequence learning itself comes from two studies that manipulated the duration of the RSI by using probabilistic target sequences

(Norman et al., 2007; Shanks et al., 2003). In Normann et al. (2007), authors first created two SOC sequences, SOC-A and SOC-B. In order to make these sequences probabilistic, the authors simply manipulated the probability that the position of the stimulus at each trial was determined by the first or second sequence. During the training phase, the position of the stimulus was determined by the SOC-A sequence in 88% of the trials (high probability trials) and by the SOC-B sequence in 12% of the trials (low probability trials). The training phase was followed by a transfer block where the probabilities were reversed (i.e., in the transfer block, the probability that a trial was determined by the SOC-A sequence was .12 and by the SOC-B sequence was .88). The authors compared two conditions where the RSI lasted either 0 ms or 1000 ms. RSI duration did not affect the magnitude of the transfer effect, which was present at both 0 ms and 1000 ms. However, the results showed that in the training blocks, participants were slower for the low probability trials compared to the high probability trials and that this effect interacted with the RSI: the RTs difference between probable and improbable trials was bigger when the RSI lasted 0 ms than when it lasted 1000 ms. Similar results were reported by Shanks (2003). Thus, short RSIs exacerbated RTs differences between probable and improbable trials compared to long RSIs, although the transfer effect was identical in both cases. Normann et al. (2007) proposed that this effect could be due to a rapid shift of attention towards the next position predicted by the sequence. When participants have more time to prepare for the next trial (in the 1000 ms condition), they would be able to expand their attentional focus in anticipation of a target appearing anywhere. This broadening of attention would reduce the detrimental effect of a stimulus arriving at an unexpected position, resulting in smaller differences between probable and improbable trials compared to the 0 ms condition.

This hypothesis could also account for the common observation that longer RSIs give rise to faster RTs (Destrebecqz & Cleeremans, 2001, 2003; Frensch & Miner, 1994; Huang et al., 2017; Norman et al., 2006, 2007; Shanks et al., 2003; Soetens et al., 1985, 2004). As the duration of the RSI increases, participants have more time to prepare for the next trial, resulting in faster RTs. Some studies have also reported that the duration of the RSI affects the slopes of RTs in the training phase: the longer the RSI, the less steep the slopes (Destrebecqz & Cleeremans, 2003; Frensch & Miner, 1994; Soetens et al., 2004; Willingham et al., 1997). This may be due to the fact that when participants are slower (due to a short RSI), there is more room for RTs to decrease. Note however that the effect of RSI on slopes has not always been found (Destrebecqz & Cleeremans, 2001; Huang et al., 2017; Norman et al., 2006, 2007; Shanks et al., 2003). Therefore, RSI duration may not affect the learning of the sequence *per se*, but the preparation of the response in the SRT task.

Hypothesis 3: RSI duration affects sequence awareness

Many studies have put forward the hypothesis that the duration of the RSI would influence the ability to elaborate a conscious representation of the knowledge acquired in an SRT task (Cleeremans & Sarrazin, 2007; Destrebecqz & Cleeremans, 2001, 2003; Frensch & Miner, 1994; Huang et al., 2017; Kuhn & Dienes, 2006; Norman et al., 2007; Savalia et al., 2016; Soetens et al., 2004; Willingham et al., 1997). According to the Awarness hypothesis, participants could only acquire explicit knowledge of the sequence when the duration of the RSI is long enough. When the RSI is very short or absent, the knowledge acquired would remain largely implicit. It is assumed that implicit knowledge tends to be restricted to perceptual features of stimuli while explicit knowledge, on the other hand, is typically associated with more abstract representations (Cleeremans & Jiménez, 2002; Huang et al., 2017). As a consequence, RSI duration would affect the degree of abstraction of the acquired knowledge.

Destrebecqz and Cleeremans (2001, 2003) evaluated the influence of the duration of the RSI on sequence awareness by adding three tasks after the SRT task. In the *fragment recognition task*, participants had to judge if sequence fragments were identical or not to the target sequence. The underlying processes that allow success in this task, i.e., retrieving in memory the sequence learned during the SRT task and comparing it to the presented fragment, are seen as largely explicit (Perruchet et al., 1997; Perruchet & Amorim, 1992; Shanks et al., 2003; Shanks & Johnstone, 1999). In the *exclusion task*, participants had to produce a different sequence from the one previously learned. The exclusion task thus requires inhibiting the sequence learned, the underlying reasoning being that this inhibition can only take place if the representation of the sequence is explicit. Finally, in the *generation task*, participants had to reproduce the target sequence in a loop. Unlike the fragment recognition task and the exclusion task, success in the generation task does not require a conscious elaboration process and can be achieved on the basis of largely implicit knowledge (but note that the question of whether the generation task require only implicit knowledge has been debated, Goschke, 1998). Destrebecqz and Cleeremans (2001, 2003) observed poorer performance with a 0 ms RSI compared to longer RSIs in the two tasks that required explicit sequence knowledge, but no effect of RSI in the more implicit task. Although the studies by Destrebecqz and Cleeremans (2001, 2003) have been criticized for their lack of power (Wilkinson & Shanks, 2004), a number of other studies using alternative measures of awareness converge on the idea that RSI affects sequence awareness (see Forkstam & Petersson, 2005; Schwarb & Schumacher, 2012 for reviews).

If the duration of the RSI affects the degree of awareness of the sequence, then the degree of abstraction of the acquired knowledge should vary accordingly. The longer the RSI duration, the more abstract the acquired knowledge will be, as it provides more time for participants to process and analyze the sequence. Therefore, the learning of abstract structural rules should be impossible at short RSIs because the knowledge of the sequence is implicit. Huang et al., (2017) investigated whether the RSI influences the ability to acquire the underlying regularity of SOC sequences, which is that the position of the stimulus at trial t is determined by the position of the stimulus at trial t_{-1} and t_{-2}. To do so, they created two SOC sequences, SOC-A and SOC-B, which shared the same higher-order structural rule (i.e., they were both SOC sequences) but whose actual realization was different. In the first 10 blocks, participants were exposed to the SOC-A sequence. Block 11 was a transfer block where the order of the stimuli was randomized. In blocks 12, 13 and 14, the order of the stimuli followed the SOC-B sequence and block 15 was a second transfer block. The authors compared participants' performance when the RSI lasted 250 ms or 750 ms. The results showed that in the first transfer block, the transfer effect was equivalent for 250 ms and 750 ms RSIs conditions.

However, in the second transfer block, only participants in the 750 ms RSI condition showed a transfer effect. Crucially, there was no transfer effect at either 250 ms or 750 ms in the control condition where participants were exposed to the SOC-B sequence for only 3 blocks, suggesting that the lack of transfer effect for SOC-B at 250 ms was not due to a lack of exposure. The authors' interpretation was that during the processing of the SOC-A sequence, participants in the 750 ms condition acquired the higher-order structural rule and reused it in the processing of the SOC-B sequence, whereas participants in the 250 ms condition failed to acquire the rule because the RSI was too short. These results are in line with the hypothesis that RSI does not affect learning as such, but the type of information that can be acquired.

4.2 Present study

In sum, both the Decay hypothesis and the Awareness hypothesis claim that RSI duration affects learning directly. According to the Decay hypothesis, RSI duration affects the amount of information that can be stored in working memory, such that shorter RSIs allow encoding more units from the sequence and therefore better sequence learning. According to the Awareness hypothesis, RSI duration affects the type of knowledge that can be acquired. Longer RSIs allow developing more abstract, higher-order properties of the sequence, and therefore better learning. In contrast to these two hypotheses, the Preparation hypothesis assumes that RSI duration does not affect sequence learning itself but participants' response preparation. Longer RSIs would reduce the surprise effect induced by a stimulus appearing at an unexpected position, thus hiding a potential transfer effect that would otherwise be observable at shorter RSIs.

Since the data reported in the literature on the impact of RSI on the transfer effect are sometimes inconsistent, and since there is no consensus on the interpretation of this impact, it may be relevant to switch to a different method to get fresh insight about the role of RSI in the SRT task. One important limitation of the transfer effect is that it is a relative measure of learning, since what is being measured is the participant's reaction to a change in the input. A slowdown in the transfer block is classically interpreted as due to the fact that the target sequence has been learned. However, because the transfer effect reflects the divergence between the target sequence and an alternative sequence, the slowdown entirely relies on the properties of this alternative sequence and how it differs from the target. Beyond the fact that this renders comparison among experiments difficult (since they vary on both the target and the alternative sequences), this method also fails to quantify *how much* participants have learned about the target sequence. Moreover, comparing averages of entire blocks does not take into account the fact that participants may continue to learn during the transfer blocks. It is therefore possible that intra-block learning hides the slowdown due to changes in the input : performance may be slower in the initial trials of the transfer block, and then improve such that the initial slowdown disappears in the average.

In the present study, we addressed these drawbacks of the transfer method. To do so, we took advantage of the properties of sequences generated by the Fibonacci grammar (Fib henceforth) which we used as a target sequence. The particularity of this sequence is that it allows us to quantify learning, trial by trial, without having to compare the performance of the participants to an alternative sequence. This provides us with a continuous measure of learning, without the need of transfer blocks. This is made possible by the fact that the regularities in the Fib sequences are dependent on one another: the learning of higher-order regularities is conditioned by the prior learning of lower order regularities. It is therefore possible to estimate how much participants have learned about the sequence by looking at the level at which these regularities have been extracted.

The Fib grammar is derived from the Lindenmayer formalism and was originally used to model algae growth (Lindenmayer, 1968; Vitányi & Walker, 1978). Recent studies have explored the processing of the Fib grammar in the SRT task (Schmid et al., 2023a; Vender et al., 2019, 2020). These studies investigated whether participants process this sequence as a recursive nested structure of events. In particular, we observed in a previous study (Schmid et al., 2023a) that the representation resulting from the processing of this sequence is similar to the natural constituent structure of the grammar. We proposed that in order to access this structure, the cognitive system would recursively merge the transitional probabilities between units of the sequence. This simple mechanism would result in a constituent structure similar to that of Fib because of the specific distribution of units in the sequence, which is aperiodic and self-similar. The Fib grammar is shown below and consists of two rewrite rules and contains a two-symbol alphabet:

$$0 \rightarrow 1$$

$$1 \rightarrow 01$$

The interpretation of these rules is the following: “0” is rewritten as “1” and “1” is rewritten as “01”. The successive application of these rules generates increasingly long sequences of 0s and 1s (henceforth refer as points). The name of this grammar comes from the fact that the number of points at each generation (i.e., each application of the rules) follows the Fibonacci sequence (Fig. 1C). This results in an asymmetry in the distribution of 0s and 1s that approximates the golden ratio (1.618) : in each generation, there are 1.618 times more 1s than 0s. Moreover, because the rewrite rules are recursive, each generation is the concatenation of the two previous ones. A sequence generated by the Fib grammar can therefore be parsed into smaller previous generations which are therefore the natural constituents of the grammar.

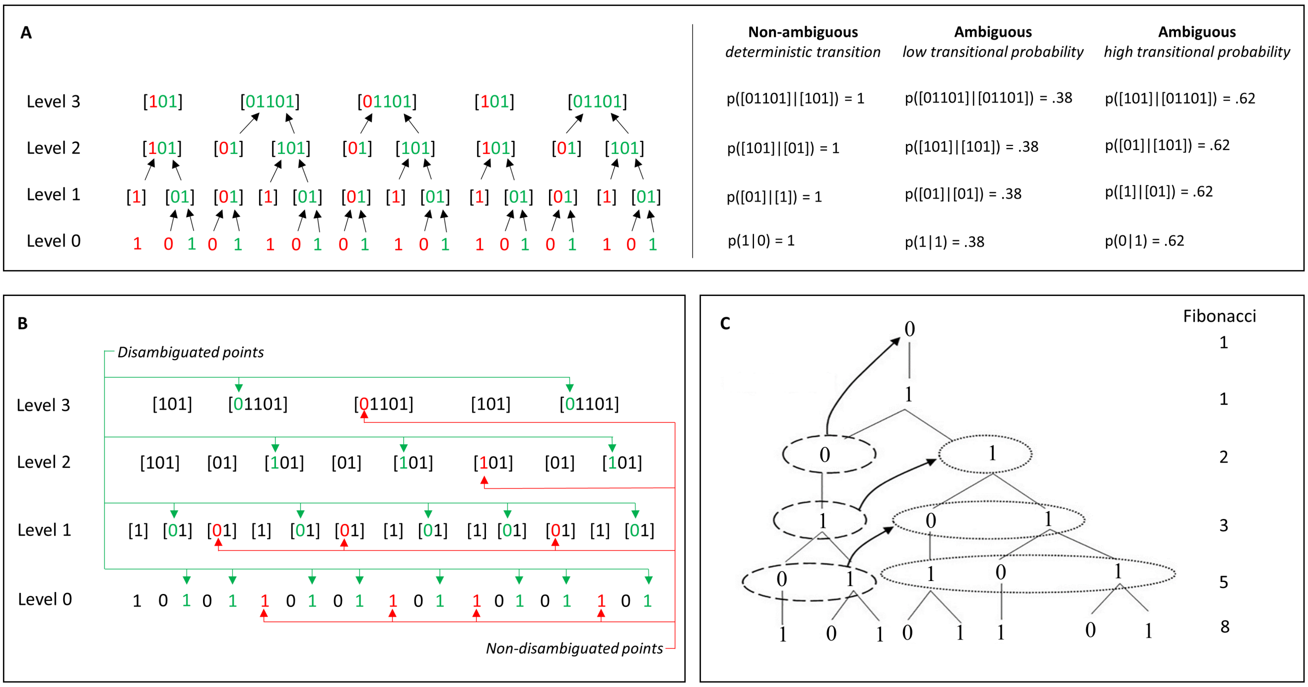


Fig. 1. (A) Left panel: depiction of the first three hierarchical levels of generation 7 of the Fibonacci grammar. Non-disambiguated points at each level are highlighted in red and disambiguated points in green. To form a new hierarchical level, points that span across a deterministic transition are combined together (this is illustrated by the arrows). The result is a new representation of the string that consists in the combination of points corresponding to natural higher-order constituents of the grammar (illustrated by the brackets). At each level, constituents spanning a deterministic transition can be combined to form an embedded hierarchy. Right panel: transition probabilities between constituents at each level. (B) Disambiguated points (green) and non-disambiguated points (red) for each hierarchical level for generation 7 of the Fibonacci grammar. In the present study, we used generation 12 of the Fibonacci grammar that consists in 233 points. We did not illustrate this generation due space limitation, but the rationale is identical. (C) Derivation of the Fibonacci grammar for the first 5 generations. The right column shows the number of symbols at each generation, which maps the Fibonacci sequence. Arrows and circles highlight the hierarchical constituency of the grammar.

The interest of the sequences generated by this grammar is that they are aperiodic and self-similar. In the classical SRT task, a SOC sequence is presented in a loop, so the sequence the participant is exposed to is periodic. Thus, learning the sequence theoretically allows to predict *all* future trials with *certainty*. This is not possible in sequences generated by the Fib grammar because of their aperiodicity: it is impossible to predict *all* the trials with *certainty* because there is no pattern that repeats in a loop. However, these sequences are not random: they present regularities, but the distribution of these regularities is aperiodic. If we examine the first order transitional probabilities (i.e., the conditional probability of a point according to the point preceding it) of these sequences, we see that three transitions are possible (Fig. 1A right panel). The first transition is deterministic: a

0 is always followed by a 1 ($p(1|0)=1$). The two other transitions are probabilistic: a 1 is followed by a 0 in 62% of the cases ($p(0|1)= .62$) and by a 1 in 38% of the cases ($p(1|1)= .38$). Points that follow a first order deterministic transition (i.e., 1s that appear after a 0) can be predicted with certainty on the basis of what precedes them, whereas this is not the case for points that follow a probabilistic transition. However, sequences generated by the Fib grammar are also self-similar, which means that the transitional probabilities between points are found also in the transitions between constituents. This implies that some of the points that follow a probabilistic transition can appear in a constituent that follows a deterministic transition. Thus, accessing to these higher-order deterministic transitions allows to predict with certainty a subset of points that follow a lower-order probabilistic transition (Fig. 1A left panel).

In a previous study (Schmid et al., 2023a), we proposed that in order to access the constituent structure of the grammar, the cognitive system would start by merging the points linked by a deterministic transition. This would result in the creation of a constituent on the basis of which the cognitive system could further detect new higher-order deterministic transitions (i.e., a deterministic transitions between constituents).

Let's take an example: at the surface level, 0 is always followed by 1; the merge of these points gives rise to the constituent [01]. This results in a new representation (which we call a new hierarchical level) where the sequence is partitioned into two constituents: [1] and [01]. At this level, constituent [01] can be followed either by constituent [1] ($p([1]\|[01])=.62$) or by constituent [01] ($p([01]\|[01])=.38$), while constituent [1] is always followed by constituent [01] ($p([01]\|[1])=1$). The cognitive system can merge again the constituents that span across a deterministic transition, which results in the creation of the constituent [101]. The new representation of the sequence then consists of two new constituents: [01] and [101]. The sequence being self-similar, the transition between these constituents is identical to that of lower levels, and merging of the deterministic transition $p([101]\|[01]) = 1$ would lead to the creation of a new hierarchical level. The key property to understand is that the first point of a constituent that follows a deterministic transition at level n always follows a probabilistic transition at level *n-1*. Thus, a point that is ambiguous at level *n* can be disambiguated if it appears at level *n+1* in a constituent that follows a deterministic transition. For example, the first point of the constituent [01] (i.e., the 0) at level 1 always follows a probabilistic transition at level 0 ($p(0|1)=.62$). If the cognitive system has detected the higher-order deterministic transition $p([01]|[1]) = 1$, then a subset of the points that follow a probabilistic transition at level 0 (i.e., the 0s that appear in constituent [01] when it follows constituent [1]) can now be predicted with certainty (Fig. 1B). For clarity, we will call *disambiguated points* the points that follow a (higher-order) deterministic transition and *non-disambiguated* points the points that follow a (higher-order) probabilistic transition at the same hierarchical level. To test the hypothesis that participants develop a hierarchical structure based on the recursive merge of deterministic transitions, we implemented sequences of the Fib grammar in an SRT task where 0s and 1s were transformed into red and blue circles and presented in the center of a screen (Schmid et al., 2023a). The RSI lasted 500 ms. The results showed a greater decrease in RTs for disambiguated points than for non-disambiguated points at levels 0, 1, 2 and 3, suggesting that participants had reached the 3rd hierarchical level.

In the present study, we asked to what extent RSI duration impacts learning in the Fib grammar. Because of the self-similar character of the sequences, the number of hierarchical levels is theoretically infinite[1], thus, there is no a priori limitation in the amount of knowledge that participants can acquire. The use of the Fib grammar therefore makes it possible to evaluate the depth of learning without having to use the transfer method.

In order to systematically explore the influence of the RSI in the SRT task, we conducted three experiments where we manipulated the duration of the RSI. The RSI lasted 1000 ms in Experiment 1, 250 ms in Experiment 2 and 100 ms in Experiment 3. In all experiments, we used the same paradigm as Schmid et al. (2023a) where sequences of the Fib grammar were implemented in the

1 Note that the hierarchical depth can of course only be infinite for an infinite string. In the present study, the shortest sequences were 144 points long and potentially involved up to 11 hierarchical levels, which is likely well beyond the processing capacity of the cognitive system.

SRT task. The 0s and 1s were transformed into red or blue circles (respectively) and presented sequentially in the center of a screen. Participants had to press the button associated with the displayed color. The answer made the circle disappear and triggered the next trial.

Each hypothesis makes distinct predictions about how RSI duration will affect the height of the hierarchical structure elaborated by the participants. According to the Decay hypothesis, the WM representation of the points/constituents deteriorates over time; this should make it more difficult to merge deterministic transitions. Therefore, the height of the hierarchical structure built by the participants should increase with the shortening of the RSI. According to the Awareness hypothesis, the duration of the RSI affects the type of knowledge that can be acquired. Under the common assumption in this field that the construction of abstract, hierarchical structure requires explicit knowledge, the height of the hierarchical structure should decrease with the shortening of the RSI. The predictions of the Awareness hypothesis regarding the height of the hierarchical structure are thus the opposite of those of the Decay hypothesis. Finally, according to the Preparation hypothesis, RSI duration does not affect learning per se but the preparation of the response in the SRT task. Thus, the height of the hierarchical structure should not vary with RSI duration. However, the length of the RSI should still affect the results because it affects the time to prepare for the next trial. Therefore, participants should be faster on average the longer the RSI. This should go along with less steep RTs slopes because if participants are faster at long RSIs, there is less room for improvement.

In order to test if RSI duration affects hierarchical learning, we conducted a first analysis in which we evaluated the height of the hierarchical structure in each experiment in the same way as in our previous study (Schmid et al., 2023a). Hierarchical elaboration generates expectations about the structure of the input, which the participants' RTs reflect (Huettel et al., 2002; Hyman, 1953; Lynn et al., 2020; McCarthy & Donchin, 1981; Sternberg, 1969). Hierarchical learning therefore manifests in terms of steeper slopes of RTs for disambiguated points at a given level compared to the slopes of non-disambiguated points at the same level. To control for asymmetry in the distribution of 0s and 1s in the sequence (i.e., 1s are more frequent than 0s), we compared at each level only 1s to 1s and 0s to 0s. Since the Preparation hypothesis predicts that RSI duration should not affect the height of the hierarchical structure but the time to prepare for the next trial, we also conducted a second analysis where we compared the average RTs and slopes of each experiment. An important point to clarify is that the Preparation hypothesis makes predictions about the average RTs and slopes for all points regardless of their ambiguity status. That's because RSI duration is a constant that affects all trials in the same way. We therefore considered disambiguated and non-disambiguated points of all levels jointly in this analysis. In order to have a wider range of RSI durations, we also included in this analysis the results of Schmid et al. (2023a) with a 500 ms RSI.

Anticipating the results, the first analysis revealed that the duration of the RSI affected the height of the hierarchical structure in a non-linear way: participants reached the 3rd hierarchical level when the RSI lasted 250 ms, whereas they only reached the 2nd hierarchical level for 100 ms and 1000 ms RSIs. None of the three hypotheses can account for this U-shape pattern of results. The second analysis partially met the predictions of the Preparation hypothesis. Concerning the average RTs, participants were slower with a 100 ms RSI compared to all other RSI durations, however, there was no difference between RSIs of 250 ms, 500 ms, 1000 ms. Concerning the average slopes, we found again a non-linear, U-shape effect of RSI duration: the slopes for the 250 ms and 500 ms RSIs were both steeper than the slopes for the 1000 ms and 100 ms RSIs, which did not differ. This result cannot be explained by the Preparation hypothesis.

In summary, the continuous measure of performance used in the present study combined with the testing of multiple RSIs show that that there is actually an optimal time window for learning in the SRT task. None of the hypotheses proposed in the literature can capture the non-linear effect of RSI duration on both the height of the hierarchical structure represented and the evolution of RTs through the task. We propose that this effect is due to an information compression mechanism determined by the interaction between the encoding capacity of the cognitive system (which corresponds to the amount of entropy the system can encode per unit of time) and the source rate of information transmission (the amount of entropy per unit of time sent by the source) (Radulescu et al., 2019, 2021; Shannon, 1948).

4.3 Experiment 1 : 1000 ms RSI

In Experiment 1, the RSI lasted 1000 ms. To avoid that the RSI of 1000 ms exhausted the participants, we made small design change in Experiment 1 compare to Schmid et al. (2023a) design (where the RSI lasted 500 ms) : we presented more blocks (7 instead of 5) but shorter (144 trials per block instead of 233). As a result, the overall number of trials per participant slightly differs from Schmid et al. (2023a) (1008 trials instead of 1165).

4.3.1 Methods

4.3.1.1 Participants

One hundred and eighty participants (49 men and 131 women; mean age 24 years old) recruited through announcements at the University of Geneva participated in the experiment. Thirty-two participated as volunteers and the remaining 148 were paid 10 CHF. All participants reported normal or corrected-to-normal vision.

4.3.1.2 Materials

The training sequence was composed of two elements and had a length of 50. The order was pseudo-randomized and elements had the same frequency. The training sequence included multiple non-grammatical sub-sequences such as 00 or 111. The longest Fib-grammatical sub-sequence had a length of 6. In the experimental blocks, the sequence consisted of generation 11 of the Fibonacci grammar which has 144 points. Each experimental block corresponded to a full generation.

4.3.1.3 Design and procedure

Each trial consisted of a red or blue circle 100px in diameter presented at the center of the screen which correspond, respectively, to 0 and 1 in a string generated by the Fib grammar. The circles disappeared after the response of the participant, or after 1200 ms, if no response was given. The RSI lasted 1000 ms. Participants responded by pressing the button corresponding to the color of the circle. Participants responded using the X and M keys of a QWERTZ keyboard (X=blue, M=red). The experiment started with a training block of 50 trials that was identical for all the participants. After the training block, participants did 7 experimental blocks of 144 trials. Instructions were displayed on the screen and participants had to click on a button to start the experiment. Participants were instructed to respond as quickly as possible. Pre-testing showed that the error rate in the task was extremely low, which is not surprising given the simplicity of the task, so the emphasis on speed alone was intended to increase the error rate and avoid ceiling effects. No information related to the grammar was given. Between each block, a message was displayed saying that participant had to press the key "enter" to start the next block, participants were told at the beginning of the experiment that they could take as much time as they wanted between each block. Stimuli were presented electronically using the E-Prime 3.0 software (Psychology Software Tools, Pittsburgh, PA). The experiment was conducted using a desktop computer running on windows 7 with a 17'' inch screen with a 1280*1020 pixels resolution. The computer screen was placed approximately 60cm from the participants. The experiment lasted approximately 30 minutes.

4.3.1.4 Data analyses

We removed six participants who had an error rate superior to 3 *SD* to the mean error rate in at least one block. Reaction times and accuracy were both modelled as dependent variables. We removed from the analysis all the trials where participants did not respond after 1200 ms (321 trials). For the analysis of reaction times, only trials with a correct answer were included. Homoscedasticity and normality were checked by visual inspection of residual plots. Data from the remaining 174 participants were analyzed with linear mixed-effects models as implemented in the lme4 package for R (Bates et al., 2014; R Core Team, 2022). Models included two fixed-effect factors and their interaction: *Exposure*, *Ambiguity,* and *Exposure*Ambiguity*. Exposure was treated as a continuous variable with a value of 0 for trials of the 1st experimental block, and of 1, 2, 3, 4, 5 and 6 for trials of the 2nd, 3d, 4th , 5th 6th and 7th blocks. This factor being continuous, it allowed us to have only one estimate which represents the evolution (i.e., the slope) of performance throughout the experiment across all participants. Ambiguity is a discrete variable contrasting disambiguated and non-disambiguated points and operationalized differently depending on the level at which its effect is explored (it is labeled *Ambiguity level$_n$* according to the level at which it has been operationalized). We entered as fixed effects the factors *Ambiguity level$_n$* (Disambiguated vs. Non-disambiguated), *Exposure,* and the interaction *Exposure*Ambiguity*. The modality "Non-disambiguated" of the factor *Ambiguity level$_n$* was always set as the intercept of the models. As random effects, the models had intercepts for *Participants*. P-values were calculated by way of the Satterthwaites's approximation to degrees of freedom with the lmerTest package (Kuznetsova et al., 2015).

In order to explore the height of the hierarchical structure elaborated by the participants, we tested the effect of Ambiguity starting at level 0. We stopped the analysis as soon as the effect was no longer significant. In the present experiment, we conducted the analysis at levels 0, 1, 2 and 3.

4.3.2 Results

Processing of Level 0

Analyses of reaction times showed a main effect of *Exposure* (β = -11.94, *SE* = 0.17, *t* = -69.06, *p* < .001) with a mean reduction of reaction times of 72 ms from block 1 to block 7. There was also a main effect of *Ambiguity level$_0$* (β = -50.23, *SE* = 0.72, *t* = -69.64, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 50 ms. The interaction *Ambiguity level$_0$* Exposure* was also significant (β = -9.59, *SE* = 0.36, *t* = -26.73, *p* < .001) with a more important reduction over time for disambiguated points ($M_{block7 - block1}$ = -95 ms) than non-disambiguated points ($M_{block7 - block1}$ = -34 ms) ($M_{block7 - block1}$ indicates the mean difference between blocks 7 and 1). Results are shown in Fig. 2.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.07, *SE* = 0.01, *z* = -7.303, *p*< .001) with a mean reduction of accuracy of 1 % from block 1 to block 7. There was also a main effect of *Ambiguity level$_0$* (β = 2.18, *SE* = 0.05, z = 47.849, p< .001) with higher accuracy for disambiguated points (M = 0.99) than for non-disambiguated points (M = 0.93). The effect of *Exposure* significantly interacted with *Ambiguity level$_0$* (β = 0.27, *SE* = 0.02, z = 11.66, p< .001) with accuracy increasing for disambiguated points over exposure ($M_{block7\ -\ block1}$ = 0.01) and decreasing for non-disambiguated points ($M_{block7\ -\ block1}$ = -0.04). Results are shown in Table 1.

Processing of Level 1

Analyses of reaction times showed a main effect of *Exposure* (β = -9.64, *SE* = 0.22, t = -44.05, p < .001) with a mean reduction of reaction times of 58 ms from block 1 to block 7. There was also a main effect of *Ambiguity level$_1$* (β = -35.82, *SE* = 0.90, t = -39.62, p < .001) with disambiguated points being faster than non-disambiguated ones by 36 ms. The interaction *Ambiguity level$_1$* Exposure* was also significant (β = -7.05, *SE* = 0.45, t = -15.65, p < .001) with a more important reduction over time for disambiguated points ($M_{block7\ -\ block1}$ = -73 ms) than non-disambiguated points ($M_{block7\ -\ block1}$ = -31 ms). Results are shown in Fig. 2.

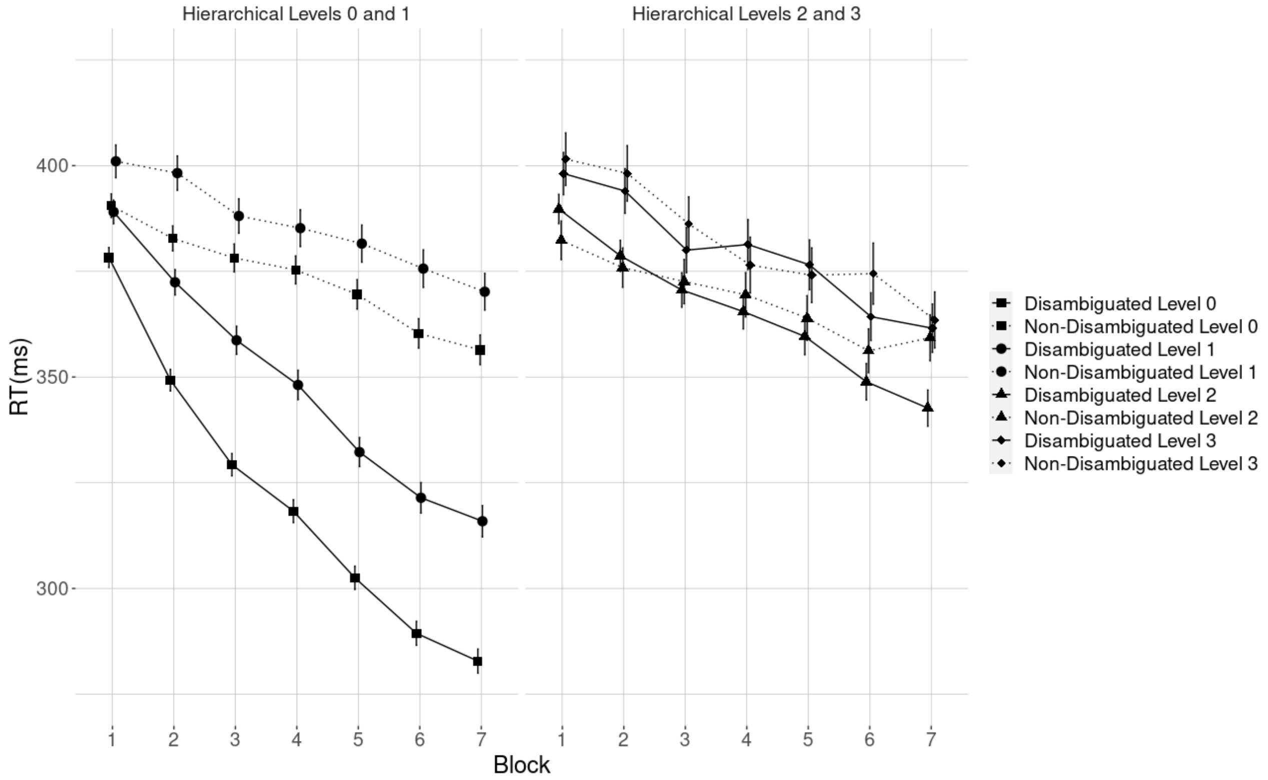


Fig. 2. *Mean RT (ms) for Disambiguated and Non-disambiguated points for Hierarchical Levels 0 and 1 (left) and for Hierarchical Levels 2 and 3 (right) by Block in Experiment 1. Errors bars denote the 95% confidence interval.*

Concerning accuracy, we found a main effect of *Exposure* (β = -0.09, *SE* = 0.01, *z* = -9.196, *p*< .001) with a mean reduction of accuracy of 2.4 % from block 1 to block 7. There was also a main effect of *Ambiguity level*$_1$ (β = 1.06, *SE* = 0.04, *z* = 27.059, *p*< .001) with accuracy higher for disambiguated points (*M* = 0.97) than for non-disambiguated points (*M* = 0.93). The effect of *Exposure* significantly interacted with *Ambiguity level*$_1$ (β = 0.10, *SE* = 0.02, *z* = 5.005, *p*< .001) with accuracy increasing for disambiguated points over time ($M_{block7 - block1}$ = 0.003) and decreasing for non-disambiguated points ($M_{block7 - block1}$ = -0.06). Results are shown in Table 1.

Table 1

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Disambiguated and Non-Disambiguated Points by Hierarchical Levels and Blocks in Experiment 1.

		Block 1		Block 2		Block 3		Block 4		Block 5		Block 6		Block 7	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 0	Disambiguated	0.98	0.13	0.99	0.11	0.99	0.08	0.99	0.08	0.99	0.09	0.99	0.08	0.99	0.08
	Non-disambiguated	0.96	0.20	0.94	0.23	0.93	0.25	0.92	0.26	0.92	0.28	0.91	0.28	0.92	0.28
Level 1	Disambiguated	0.97	0.16	0.98	0.15	0.97	0.16	0.97	0.18	0.97	0.17	0.97	0.17	0.97	0.16
	Non-disambiguated	0.96	0.20	0.94	0.23	0.93	0.26	0.92	0.27	0.92	0.27	0.91	0.28	0.90	0.30
Level 2	Disambiguated	0.95	0.21	0.94	0.23	0.94	0.25	0.93	0.26	0.92	0.27	0.92	0.28	0.92	0.27
	Non-disambiguated	0.97	0.19	0.94	0.23	0.93	0.25	0.92	0.28	0.91	0.28	0.91	0.29	0.91	0.28
Level 3	Disambiguated	0.96	0.20	0.94	0.24	0.92	0.27	0.92	0.27	0.92	0.27	0.92	0.28	0.90	0.30
	Non-disambiguated	0.96	0.19	0.95	0.22	0.93	0.25	0.93	0.26	0.92	0.27	0.92	0.28	0.91	0.29

Processing of Level 2

Analyses of reaction times showed a main effect of *Exposure* (β = -5.73, *SE* = 0.27, t = -20.893, p < .001) with a mean reduction of reaction times of 34 ms from block 1 to block 7. There was also a main effect of *Ambiguity* $level_2$ (β = -5.07, *SE* = 1.14, t = -4.436, p < .001) with disambiguated points being faster than non-disambiguated ones by 5 ms. The interaction *Ambiguity* $level_2$*Exposure* was also significant (β = -3.85, *SE* = 0.57, t = -6.762, p < .001) with a more important reduction over exposure for disambiguated points ($M_{block7\,-\,block1}$ = -47 ms) than non-disambiguated points ($M_{block7\,-\,block1}$ = -23 ms). Results are shown in Fig. 2.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.12, *SE* = 0.01, z = -11.66, p< .001) with a mean reduction of accuracy of 4 % from block 1 to block 7. There was no effect of *Ambiguity* $level_2$ (β = 0.05, *SE* = 0.04, z = 1.33, p = .184) and the interaction *Ambiguity* $level_2$*Exposure* was not significant (β = 0.04, *SE* = 0.02, z = 1.852, p= .064). Results are shown in Table 1.

Processing of Level 3

Analyses of reaction times showed a main effect of *Exposure* (β = -5.22, *SE* = 0.37, *t* = -14.23, *p* < .001) with a mean reduction of reaction times of 31 ms from block 1 to block 7. There was no effect of *Ambiguity level$_3$* (β = -1.94, *SE* = 1.50, *t* = -1.293, *p* = .196). The interaction *Ambiguity level$_3$*Exposure* was also not significant (β = -0.03, *SE* = 0.7, *t* = -0.042, *p* = .966). Results are shown in Fig. 2.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.12, *SE* = 0.01, *z* = -9.864, *p*< .001) with a mean reduction of accuracy of 6 % from block 1 to block 7. There was also a main effect of *Ambiguity level$_3$* (β = -0.13, *SE* = 0.05, *z* = -2.622, *p* = .009) with accuracy higher for non-disambiguated points (*M* = 0.92) than for disambiguated points (*M* = 0.93). The interaction *Ambiguity level$_3$* Exposure* was however not significant (β = 0.02, *SE* = 0.03, *z* = 0.848, *p*= .396). Results are shown in Table 1.

4.3.3 Discussion

In Experiment 1, we found that RTs of disambiguated points at levels 0, 1 and 2 decreased significantly more through exposure than their non-disambiguated counterparts. We also found that accuracy tended to decreased over time at levels 0, 1, 2 and 3. However, this decrease was modulated by Ambiguity : at levels 0 and 1, accuracy for non-disambiguated points decreased through exposure while it increased for disambiguated points. At level 2, the decrease in accuracy was identical for disambiguated and non-disambiguated points. Thus, even if this decrease in accuracy suggests a speed-accuracy trade-off, it cannot explain the difference in RTs between disambiguated and non-disambiguated points. This decrease in accuracy was also observed in our previous study (Schmid et al., 2023a) and could be due to instructions that emphasized speed of response or to the boredom induced by the long RSI or the simplicity of the task. At level 3, RTs did not differ between disambiguated and non-disambiguated points. There was however a main effect in accuracy with a better performance for disambiguated points compared to non-disambiguated points. However, this effect did not interact with time, suggesting that the effect was present from the beginning and does therefore not reflect learning. Taken together, these results suggest that with an RSI of 1000 ms, participants reached the second hierarchical level.

4.4 Experiment 2 : 250 ms RSI

In Experiment 2, the RSI lasted 250 ms. Based on pilot testing, we noticed that the fatigue induced by this duration is reduced compared to that induced by the RSI of 1000 ms in Experiment 1. We therefore reproduced the design of Schmid et al. (2023a) with 5 blocks of 233 trials each. As a result, the total number of trials per participant is slightly higher in Experiment 2 (1165 trials) than in Experiment 1 (1008 trials). Apart from these minor adjustments, the design of Experiment 2 was identical to that of Experiment 1.

4.4.1 Methods

4.4.1.1 Participants

One hundred and fifty participants (men and women; mean age years old) recruited using Prolific (www.prolific.co) participated in the experiment. Participants were paid 3.75 £. All participants reported normal or corrected-to-normal vision.

4.4.1.2 Materials

The training sequence was identical to that of experiment 1. However, instead of generation 11 of the Fib grammar we used generation 12 which contain 233 points.

4.4.1.3 Design and procedure

The procedure was identical to that of Experiment 1 except for the following elements. During the training block, when the participants made an error, the experiment stopped and a message appeared to remind them the color – key association, the experiment resumed after 3000 ms. In the experimental blocks, no message appeared when they made an error. After the training block, participants did 5 experimental blocks of 233 trials. The experiment was created using PsychoPy (Peirce et al., 2019) and conducted online on the website Pavlovia (www.pavlovia.org). Participants were asked to perform the experiment in a quiet environment where they could not be disturbed. The experiment lasted approximately 20 minutes.

4.4.1.4 Data analyses

We removed one participant who had a number of timeout trials 3 SD above the mean answered trials. We also removed four participants who had an error rate superior to 3 *SD* to the mean error rate in at least one block. We removed from the analysis all the trials where participants did not respond after 1200 ms (578 trials). Data from the remaining 145 participants were analyzed in the same way as in Experiment 1.

4.4.2 Results

Processing of Level 0

Analyses of reaction times showed a main effect of *Exposure* (β = -19.81, SE = 0.24, t = -80.99, p < .001) with a mean reduction of reaction times of 87 ms from block 1 to block 5. There was also a main effect of *Ambiguity level*$_0$ (β = -51.24, SE = 0.71, t = -71.30, p < .001) with disambiguated points being faster than non-disambiguated ones by 51 ms. The interaction *Ambiguity level*$_0$* *Exposure* was also significant (β = -14.98, SE = 0.51, t = -29.64, p < .001) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -110 ms) than non-disambiguated points ($M_{block5 - block1}$ = -45 ms) ($M_{block5 - block1}$ indicates the mean difference between blocks 1 and 5). Results are shown in Fig. 3.

Concerning accuracy, there was a main effect of *Exposure* (β = 0.04, SE = 0.01, z = -2.634, p = .008) with a mean reduction of accuracy of 0.4 % from block 1 to block 5. There was also a main effect of *Ambiguity level*$_0$ (β = 1.99, SE = 0.05, z = 42.59, p < .001) with higher accuracy for disambiguated points (M = 0.99) than for non-disambiguated points (M = 0.94). The interaction

*Exposure*Ambiguity level*$_0$ was significant (β = 0.14, *SE* = 0.03, *z* = 4.294, *p < .001*) with accuracy increasing over exposure for disambiguated points ($M_{block5\,-\,block1}$ = 0.004) and decreasing for non-disambiguated points ($M_{block5\,-\,block1}$ = -0.02). Results are shown in Table 2.

Processing of Level 1

Analyses of reaction times showed a main effect of *Exposure* (β = -19.26, *SE* = 0.34, *t* = -57.22, *p* < .001) with a mean reduction of reaction times of 77 ms from block 1 to block 5. There was also a main effect of *Ambiguity level*$_1$ (β = -87.57, *SE* = 1.00, *t* = -87.86, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 88 ms. The interaction *Ambiguity level*$_1$* *Exposure* was also significant (β = -18.40, *SE* = 0.70, *t* = -26.36, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block5\,-\,block1}$ = -108 ms) than non-disambiguated points ($M_{block5\,-\,block1}$ = -30 ms). Results are shown in Fig. 3.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.12, *SE* = 0.01, *z* = -9.492, *p* < .001) with a mean reduction of accuracy of 2.4 % from block 1 to block 5. There was also a main effect of *Ambiguity level*$_1$ (β = 1.56, *SE* = 0.04, *z* = 39.80, *p*< .001) with accuracy higher for disambiguated points (*M* = 0.98) than for non-disambiguated points (*M* = 0.90). The effect of *Exposure* significantly interacted with *Ambiguity level*$_1$ (β = 0.11, *SE* = 0.03, *z* = 3.803, *p*< .001) with accuracy increasing over exposure for disambiguated points ($M_{block5\,-\,block1}$ = 0.004) and decreasing for non-disambiguated points ($M_{block1\,-\,block75}$ = -0.06). Results are shown in Table 2.

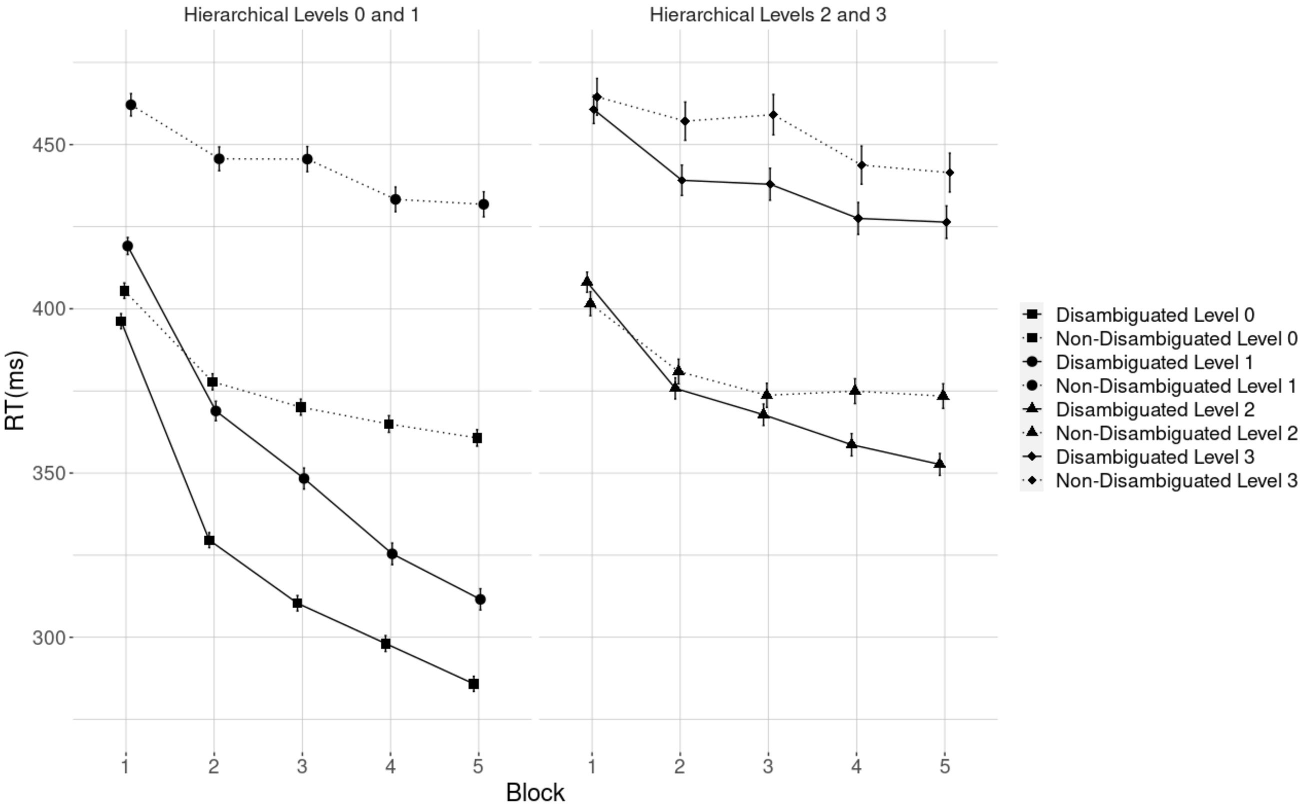


Fig. 3. *Mean RT (ms) for Disambiguated and Non-disambiguated points for Hierarchical Levels 0 and 1 (left) and Hierarchical Levels 2 and 3 (right) by Block in Experiment 2. Errors bars denote the 95% confidence interval.*

Processing of Level 2

Analyses of reaction times showed a main effect of *Exposure* (β = -10.32, *SE* = 0.34, *t* = -30.19, *p* < .001) with a mean reduction of reaction times of 41 ms from block 1 to block 5. There was also a main effect of *Ambiguity level*$_2$ (β = -7.16, *SE* = 0.99, *t* = -7.21, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 7 ms. The interaction *Ambiguity level*$_2$* *Exposure* was also significant (β = -6.33, *SE* = 0.70, *t* = -9.039, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block5\,-\,block1}$ = -55 ms) than non-disambiguated points ($M_{block5\,-\,block1}$ = -28 ms). Results are shown in Fig. 3.

Concerning accuracy, there was a significant effect of *Exposure* (β = -0.07, *SE* = 0.02, *z* = -4.316, *p < .001*) with a mean augmentation of accuracy of 1.6 % from block 1 to block 5. There was also a main effect of *Ambiguity level*$_2$ (β = -0.09, *SE* = 0.04, *z* = -2.103, *p* = .035) with accuracy higher for non-disambiguated points (*M* = 0.942) than for disambiguated points (*M* = 0.937). The effect of

Exposure significantly interacted with *Ambiguity level*$_2$ (β = 0.14, *SE* = 0.03, *z* = 4.508, *p*< .001) with accuracy decreasing more over exposure for non-disambiguated points ($M_{block5 - block1}$ = -0.03) than for disambiguated points ($M_{block5 - block1}$ = -0.005). Results are shown in Table 2.

Table 2

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Disambiguated and Non-Disambiguated Points by Hierarchical Levels and Blocks in Experiment 2.

		Block 1		Block 2		Block 3		Block 4		Block 5	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 0	Disambiguated	0.98	0.11	0.99	0.09	0.99	0.09	0.99	0.09	0.99	0.09
	Non-disambiguated	0.95	0.21	0.94	0.24	0.94	0.24	0.93	0.25	0.93	0.25
Level 1	Disambiguated	0.98	0.15	0.98	0.15	0.97	0.16	0.97	0.16	0.98	0.16
	Non-disambiguated	0.94	0.24	0.91	0.29	0.88	0.32	0.88	0.32	0.88	0.32
Level 2	Disambiguated	0.94	0.23	0.93	0.25	0.94	0.25	0.94	0.24	0.94	0.24
	Non-disambiguated	0.97	0.18	0.94	0.23	0.94	0.24	0.93	0.25	0.93	0.25
Level 3	Disambiguated	0.93	0.26	0.91	0.29	0.88	0.32	0.89	0.31	0.89	0.31
	Non-disambiguated	0.95	0.22	0.90	0.30	0.88	0.33	0.87	0.34	0.87	0.33

Processing of Level 3

Analyses of reaction times showed a main effect of *Exposure* (β = -7.34, *SE* = 0.52, *t* = -14.01, *p* < .001) with a mean reduction of reaction times of 30 ms from block 1 to block 5. There was also a main effect of *Ambiguity level*$_3$ (β = -14.38, *SE* = 1.55, *t* = -9.277, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 14 ms. The interaction *Ambiguity level*$_3$* *Exposure* was significant (β = -2.14, *SE* = 1.09, *t* = -1.967, *p* = .049) with a more important reduction over exposure for disambiguated points ($M_{block5 - block1}$ = -34 ms) than non-disambiguated points ($M_{block5 - block1}$ = -23 ms). Results are shown in Fig. 3.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.15, *SE* = 0.02, *z* = -9.951, *p*< .001) with a mean reduction of accuracy of 5.6 % from block 1 to block 5. There was no main effect of *Ambiguity level*$_3$ (β = 0.06, *SE* = 0.05, *z* = 1.296, *p* = .195). The interaction *Exposure**

Ambiguity $level_3$ was however significant (β = 0.09, *SE* = 0.03, *z* = 2.762, *p* = .006) with accuracy decreasing less over exposure for disambiguated points ($M_{block5 - block1}$ = -0.04) than for non-disambiguated points ($M_{block5 - block1}$ = -0.08). Results are shown in Table 2.

Processing of level 4

Analyses of reaction times showed a main effect of *Exposure* (β = -6.21, *SE* = 0.53, *t* = -11.716, *p* < .001) with a mean reduction of reaction times of 25 ms from block 1 to block 5. There was no effect of *Ambiguity* $level_4$ (β = -0.79, *SE* = 1.53, *t* = -0.515, *p* = .607). The interaction *Ambiguity* $level_4$* *Exposure* did not reach significance (β = -1.04, *SE* = 1.08, *t* = -0.962, *p* = .336).

Concerning accuracy, we found a main effect of *Exposure* (β = -0.15, *SE* = 0.03, *z* = -5.856, *p*< .001) with a mean reduction of accuracy of 3.4 % from block 1 to block 5. There was a main effect of *Ambiguity* $level_4$ (β = 0.15, *SE* = 0.07, *z* = 2.090, *p* = .037) with accuracy higher for disambiguated points (*M* = 0.946) than for non-disambiguated points (*M* = 0.939). The interaction *Exposure** *Ambiguity* $level_4$ was however not significant (β = 0.01, *SE* = 0.05, *z* = 0.134, *p* = .893).

4.4.3 Discussion

The results of Experiment 2 show that RTs of disambiguated points decreased significantly more through exposure than their non-disambiguated counterparts at hierarchical levels 0, 1, 2 and 3. As in Experiment 1, accuracy tended to decrease throughout the experiment, but this decrease was modulated by ambiguity: at levels 0 and 1, accuracy increased for disambiguated points, while it decreased for their non-disambiguated counterparts. At levels 2 and 3, accuracy decreased for both disambiguated and non-disambiguated points, but the decrease was significantly greater for non-disambiguated points. Finally, there was no effect at level 4 in the RTs. Accuracy turned out to be higher for non-disambiguated points than for disambiguated points, but this effect did not evolve over time and has a small effect size (less than 1%); we will therefore not interpret this effect. In summary, results of Experiment 2 suggest that participants reached the third hierarchical level when the RSI lasted 250 ms.

4.5 Experiment 3: 100 ms RSI

In Experiment 3, the RSI was 100 ms long. We took advantage of the fact that this short RSI reduces the total time to complete the experiment to increase the number of trials per participant. Thus, Experiment 3 was slightly longer but the number of trials was more than twice as large. This results in a higher statistical power compared to Experiments 1 and 2, however, as will become clear, this increase in statistical power did not favor the appearance of significant effects. Participants were exposed to 7 blocks of 377 trials (generation 13 of the Fib grammar) for a total of 2639 trials (compared to 1008 trials for experiment 1 and 1165 trials for experiment 2). Apart from these differences, the design of Experiment 3 was identical to that of Experiments 1 and 2.

4.5.1 Methods

4.5.1.1 Participants

One hundred participants (44 men and 56 women; mean age 24.4 years old) recruited using Prolific (www.prolific.co) participated in the experiment. Participants were paid 3.75 £. All participants reported normal or corrected-to-normal vision.

4.5.1.2 Materials

The training sequence was identical to that of experiment 1 and 2. However, we used generation 13 the Fib grammar which contain 377 points instead of generation 11 used in Experiment 1 and generation 12 used in Experiment 2.

4.5.1.3 Design and procedure

The procedure was identical to that of Experiment 2 except that participants did 7 experimental blocks of 377 trials. The experiment was created using PsychoPy (Peirce et al., 2019) and conducted online on the website Pavlovia (www.pavlovia.org). Participants were asked to perform the experiment in a quiet environment where they could not be disturbed. The experiment lasted approximately 30 minutes.

4.5.1.4 Data analyses

One participant was removed for not providing answers in two blocks. We also removed five participants who had an error rate superior to 3 *SD* to the mean error rate in at least one block. We removed from the analysis all the trials where participants did not respond after 1200 ms (1375 trials). For the analysis of reaction times, only trials with a correct answer were included. Homoscedasticity and normality were checked by visual inspection of residual plots. Data from the remaining 94 participants were analyzed in the same way as in Experiment 1 and 2.

4.5.2 Results

Processing of Level 0

Analyses of reaction times showed a main effect of *Exposure* (β = -11.47, *SE* = 0.17, *t* = -69.38, *p* < .001) with a mean reduction of reaction times of 69 ms from block 1 to block 7. There was also a main effect of *Ambiguity level*$_0$ (β = -53.67, *SE* = 0.68, *t* = -78.39, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 54 ms. The interaction *Ambiguity level*$_0$* *Exposure* was also significant (β = -5.23, *SE* = 0.34, *t* = -15.28, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block7-block1}$ = -91 ms) than non-disambiguated points ($M_{block7-block1}$ = -52 ms) ($M_{block7-block1}$ indicates the mean difference between blocks 1 and 7). Results are shown in Fig. 4.

Concerning accuracy, there was no effect of *Exposure* (β = 0.01, *SE* = 0.01, *z* = 1.54, *p* = .123). There was however a main effect of *Ambiguity level*$_0$ (β = 1.57, *SE* = 0.04, *z* = 42.637, *p* < .001) with higher accuracy for disambiguated points (*M* = 0.99) than for non-disambiguated points (*M* = 0.95). The interaction *Exposure**Ambiguity level*$_0$ was not significant (β = -0.02, *SE* = 0.02, *z* = -0.977, *p* = .328). Results are shown in Table 3.

Processing of Level 1

Analyses of reaction times showed a main effect of *Exposure* (β = -13.30, *SE* = 0.22, *t* = -61.07, *p* < .001) with a mean reduction of reaction times of 80 ms from block 1 to block 7. There was also a main effect of *Ambiguity level₁* (β = -141.87, *SE* = 0.91, *t* = -155.89, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 142 ms. The interaction *Ambiguity* $level_1$* *Exposure* was also significant (β = -13.38, *SE* = 0.45, *t* = -29.66, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block7 - block1}$ = -117 ms) than non-disambiguated points ($M_{block7 - block1}$ = -31 ms). Results are shown in Fig. 4.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.09, *SE* = 0.01, *z* = -11.89, *p*< .001) with a mean reduction of accuracy of 2.8 % from block 1 to block 7. There was also a main effect of *Ambiguity* $level_1$ (β = 1.67, *SE* = 0.03, *z* = 50.89, *p*< .001) with accuracy higher for disambiguated points (*M* = 0.98) than for non-disambiguated points (*M* = 0.89). The effect of *Exposure* significantly interacted with *Ambiguity* $level_1$ (β = 0.07, *SE* = 0.02, *z* = 4.431, *p*< .001) with accuracy increasing over exposure for disambiguated points ($M_{block7 - block1}$ = 0.004) and decreasing for non-disambiguated points ($M_{block7 - block1}$ = -0.07). Results are shown in Table 3.

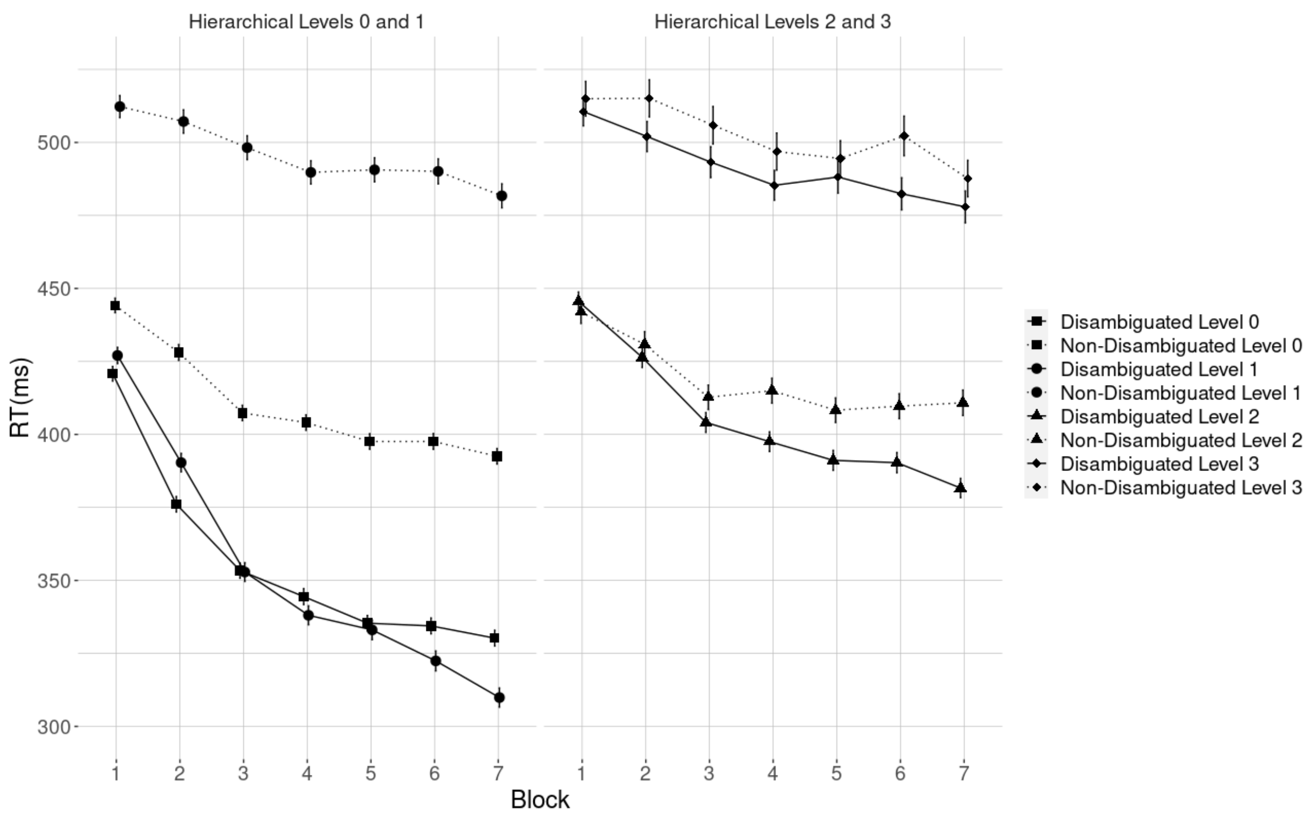


Fig. 4. *Mean RT (ms) for Disambiguated and Non-disambiguated points of Hierarchical Levels 0 and 1 (left) and for Hierarchical Levels 2 and 3 (right) by Block in Experiment 3. Errors bars denote the 95% confidence interval.*

Table 3

Mean Proportion (M) and Standard Deviation (SD) of Correct Responses for Disambiguated and Non-Disambiguated Points by Hierarchical Levels and Blocks in Experiment 3.

		Block 1		Block 2		Block 3		Block 4		Block 5		Block 6		Block 7	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Level 0	Disambiguated	0.99	0.10	0.99	0.10	0.99	0.11	0.99	0.10	0.98	0.12	0.99	0.9	0.99	0.10
	Non-disambiguated	0.95	0.22	0.94	0.23	0.95	0.22	0.96	0.21	0.95	0.21	0.95	0.22	0.95	0.21
Level 1	Disambiguated	0.98	0.14	0.98	0.15	0.98	0.15	0.98	0.15	0.97	0.17	0.98	0.15	0.97	0.16
	Non-disambiguated	0.94	0.23	0.91	0.29	0.88	0.32	0.89	0.32	0.87	0.33	0.87	0.33	0.87	0.33
Level 2	Disambiguated	0.95	0.23	0.94	0.24	0.95	0.23	0.96	0.21	0.95	0.22	0.95	0.22	0.96	0.20
	Non-disambiguated	0.96	0.19	0.95	0.21	0.96	0.21	0.96	0.21	0.96	0.20	0.95	0.22	0.94	0.22
Level 3	Disambiguated	0.94	0.24	0.90	0.30	0.88	0.33	0.90	0.31	0.87	0.34	0.88	0.33	0.88	0.33
	Non-disambiguated	0.95	0.22	0.92	0.27	0.89	0.31	0.87	0.33	0.87	0.33	0.87	0.34	0.88	0.33

Processing of Level 2

Analyses of reaction times showed a main effect of *Exposure* (β = -8.10, *SE* = 0.23, *t* = -35.56, *p* < .001) with a mean reduction of reaction times of 48 ms from block 1 to block 7. There was also a main effect of *Ambiguity* $level_2$ (β = -12.55, *SE* = 0.94, *t* = -13.36, *p* < .001) with disambiguated points being faster than non-disambiguated ones by 13 ms. The interaction *Ambiguity* $level_2$* *Exposure* was also significant (β = -4.86, *SE* = 0.47, *t* = -10.359, *p* < .001) with a more important reduction over exposure for disambiguated points ($M_{block7-block1}$ = -64 ms) than non-disambiguated points ($M_{block7-block1}$ = -31 ms). Results are shown in Fig. 4.

Concerning accuracy, there was no effect of *Exposure* (β = 0.02, *SE* = 0.01, *z* = 1.805, *p* = .071). There was a main effect of *Ambiguity* $level_2$ (β = -0.13, *SE* = 0.04, *z* = -3.136, *p* = .002) with accuracy higher for non-disambiguated points (*M* = 0.96) than for disambiguated points (*M* = 0.95). The effect of *Exposure* significantly interacted with *Ambiguity* $level_2$ (β = 0.09, *SE* = 0.02, *z* = 4.324, *p*< .001) with accuracy increasing for disambiguated points over exposure ($M_{block7-block1}$ =

0.01) and decreasing for non-disambiguated points ($M_{block7 - block1}$ = -0.02). Results are shown in Table 3.

Processing of Level 3

Analyses of reaction times showed a main effect of *Exposure* (β = -4.75, *SE* = 0.35, t = -13.449, p < .001) with a mean reduction of reaction times of 29 ms from block 1 to block 7. There was also a main effect of *Ambiguity level*$_3$ (β = -10.33, *SE* = 1.46, t = -7.091, p < .001) with disambiguated points being faster than non-disambiguated ones by 10 ms. The interaction *Ambiguity level*$_3$* *Exposure* was however not significant (β = -0.86, *SE* = 0.72, t = -1.182, p = .237). Results are shown in Fig. 4.

Concerning accuracy, we found a main effect of *Exposure* (β = -0.11, *SE* = 0.01, z = -12.31, p< .001) with a mean reduction of accuracy of 6.7 % from block 1 to block 7. There was no effect of *Ambiguity level*$_3$ (β = -0.04, *SE* = 0.04, z = -0.986, p = .324). The interaction *Exposure** *Ambiguity level*$_3$ was also not significant (β = 0.03, *SE* = 0.02, z = 1.792, p = .073). Results are shown in Table 3.

4.5.3 Discussion

The results of Experiment 3 showed that the RTs for disambiguated points decreased more through exposure than their non-disambiguated counterparts at levels 0, 1 and 2. There was no effect in accuracy at level 0. At levels 1 and 2, accuracy increased through exposure for disambiguated points and decreased for non-disambiguated points. At level 3, there was only a main effect on RTs with disambiguated points processed faster than non-disambiguated points. However, the interaction was non-significant, suggesting that this effect do not reflect learning. Taken together, these results suggest that participants reached the second hierarchical level when the RSI lasted 100 ms. It is interesting to note that even though the number of trials was significantly higher in this experiment, this greater exposure did not improve learning compared to Experiment 2 where participants reached level 3. This is in line with a recent meta-analysis (Isbilen & Christiansen, 2022), which reported that the amount of exposure does not influence learning. It may be that the amount of exposure plays only a minor role above a certain threshold.

4.6 Comparison of 1000 ms, 500 ms, 250 ms and 100 ms RSIs

In this second analysis, we test the predictions of the Preparation hypothesis that RSI duration affects participants' control of the response. With longer RSIs, participants have more time to prepare their responses which should result in faster RTs than for shorter RSI. Conversely, the slope of RTs should be steeper the shorter the RSI because there is more room for improvement. Since the RSI is the same throughout the trials, the influence of RSI duration on preparation is also constant and is therefore not expected to vary between the different types of points. We therefore compared the average RTs and slopes without distinguishing between disambiguated and non-disambiguated points at the different levels. In order to have a wider range of RSI duration, we also integrated in this analysis our previous results where the RSI lasted 500 ms (Schmid et al., 2023a). Except for the duration of the RSI, the design of this experiment was strictly identical to that of Experiment 2 where the RSI lasted 250 ms. Since the amount of exposure varied in each experiment, we considered only the first 1008 trials of each experiment in order to have the same number of trials in each experiment. This number corresponds to the number of trials in Experiment 1, which was the shortest.

4.6.1 Methods

4.6.1.1 Materials

Since Experiment 1 contains 1008 experimental trials and is the shortest, we considered in the analyses the first 1008 experimental trials of experiments 1 (RSI = 1000 ms), 2 (RSI = 250 ms) and 3 (RSI = 100 ms). We also included in the analysis the first 1008 experimental trials of Schmid et al, (2023a) where the RSI was 500 ms. We did not include in the analysis the trials of the training block where the order of the points was random.

4.6.1.2 Data analysis

Models included two fixed-effect factors and their interaction: *Exposure*, *RSI*, and *Exposure *RSI*. Exposure was treated as a continuous variable with a value of 0 for the first trial and 1007 for the last trial. This factor being continuous, it allowed us to have only one estimate which represents the evolution (i.e., the slope) of RTs throughout the experiments across all participants. Trials where an incorrect answer was given were not included in the analysis. RSI is a between subject discrete variable contrasting RSI duration. We entered as fixed effects the factors *RSI* (1000 ms vs 500 ms vs 250 ms vs 100 ms), *Exposure,* and the interaction *Exposure***RSI*. The modality "100 ms" of the factor *RSI* was set as the intercept of the models. As random effects, the models had intercepts for *Participants*. Since the factor RSI contained 4 modalities and that the Preparation hypothesis makes predictions on all comparisons, we had to run the model 3 times. In order to control for type 1 error, we applied the Bonferroni correction for multiple testing by dividing the alpha level by 3. We therefore considered as significant the p-value lower than .01667. P-values were calculated by way of the Satterthwaites's approximation to degrees of freedom with the lmerTest package (Kuznetsova et al., 2015).

4.6.2 Results

Results showed a main effect of the *Exposure* (β = -0.089, *SE* = 0.0005, t = -160.91, p < .001) indicating that RTs decreased across exposure. RTs in the "100 ms" condition (*M* = 416 ms; *SD* = 146 ms) were significantly slower than those of the "250 ms" condition (*M* = 365 ms; *SD* = 138 ms) (β = -51.87, *SE* = 8.39, t = -6.182, p < .001), "500 ms" condition (*M* = 363 ms; *SD* = 136 ms) (β = -53.54, *SE* = 8.24, t = -6.494, p < .001) and "1000 ms" condition (*M* = 350 ms; *SD* = 134 ms) (β = -67.40, *SE* = 8.11, t = -8.310, p < .001). The mean RTs of the "1000 ms" condition did not differ from those of the "500 ms" condition (β = 13.86, *SE* = 6.95, t = 1.994, p = .047) and "250 ms" condition (β = 15.53, *SE* = 7.12, t = 2.180, p = .030). Finally, there was no difference between the "500 ms" and "250 ms" conditions (β = 1.66, *SE* = 7.27, t = 0.229, p = .819). Concerning the interaction *RSI*Exposure*, there was no difference in slope of RTs between the “100 ms” condition and “1000 ms” condition (β = -0.03, *SE* = 0.001, *t* = -1.622, *p* = .105). There was also no difference between the “250 ms” and “500 ms” conditions (β = 0.002, *SE* = 0.002, *t* = 1.179, *p* = .238). However, the RTs decreased more through exposure for the "250 ms" condition compare to the “1000 ms” condition (β = -0.02, *SE* = 0.001, *t* = -13.872, $p < .001$) and the “100 ms” condition (β = -0.02, *SE* = 0.001, *t* = -10.209, $p < .001$). Finally, RTs decreased more through exposure for the "500 ms" condition compare to the “1000 ms” condition (β = -0.02, *SE* = 0.001, *t* = -15.396, $p < .001$) and the “100 ms” condition (β = -0.02, *SE* = 0.002, *t* = -11.404, $p < .001$). Results are shown in Fig. 5.

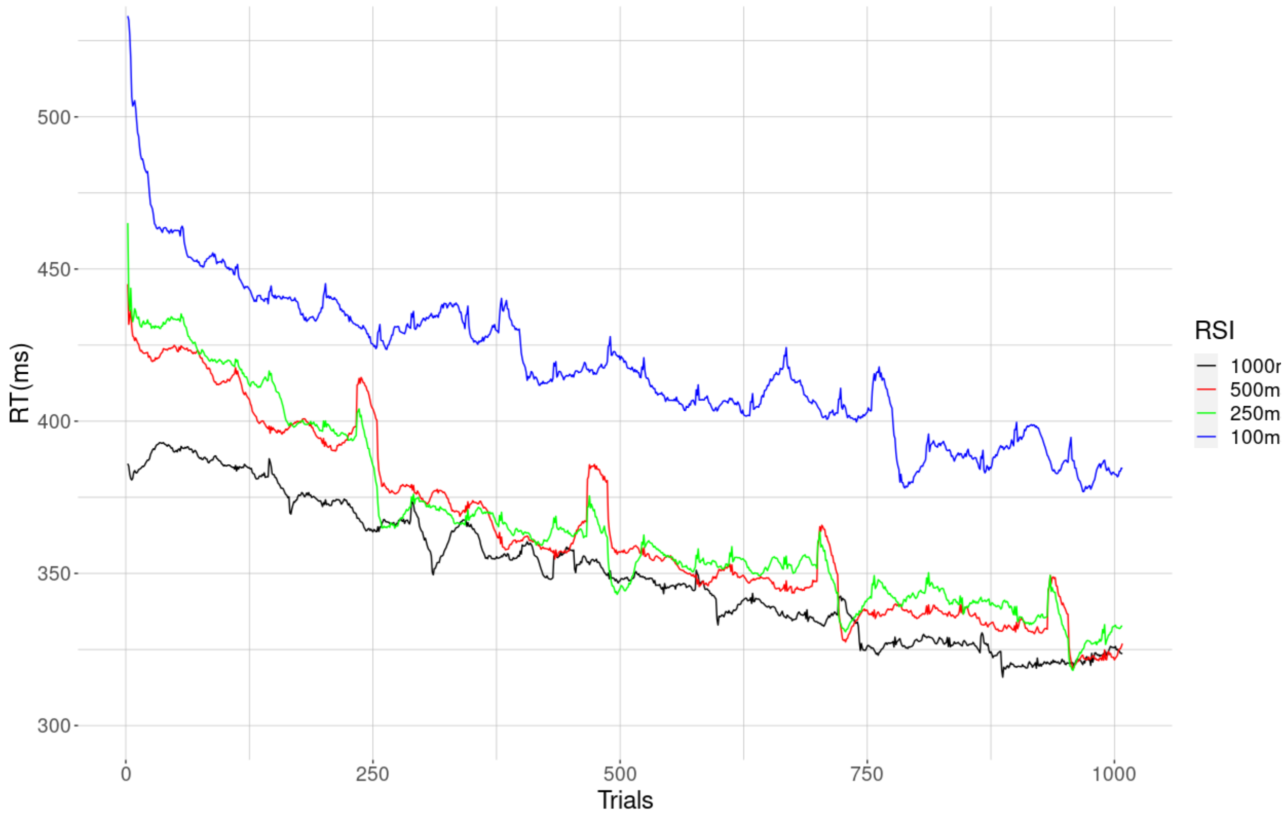


Fig. 5. *Moving average of RTs for the first 1008 experimental trials as a function of RSI duration. The moving average is calculated over the 21 trials that precede trial t. We removed trial 1 of each experiment from the graph because it was processed extremely slowly and was out of the frame.*

4.6.3 Discussion

According to the Preparation hypothesis, average RTs should increase as the RSI is reduced. In line with this prediction, participants were indeed slower when the RSI was 100 ms long compared to all other conditions. However, we found no difference in average RTs between the 1000 ms, 500 ms and 250 ms RSIs, contrary to the predictions of the Preparation hypothesis. We doubt that this lack of effect is due to a lack of power given the number of participants (572) and the number of trials per participant (1008). We also found that the duration of the RSI affected the slope of the RTs in a non-linear way: RTs decreased more strongly for RSIs of 250 ms and 500 ms compared to RSIs of 1000 ms and 100 ms. Again, this result cannot be explained by the Preparation hypothesis which predicted a linear relationship between slope steepness and RSI.

4.7. General discussion

Three hypotheses have been put forward to explain the role of RSI duration on sequence learning in the SRT task. According to the Decay hypothesis (Frensch & Miner, 1994; Soetens et al., 2004), the duration of the RSI affects information processing in WM. As the duration of the RSI increases, the trace of the stimuli would tend to decrease, which would decrease the number of stimuli simultaneously active in WM. As a result, the detection of sequence regularities would become more difficult as RSI increases and conversely, learning would be better with shorter RSIs. According to the Preparation hypothesis (Norman et al., 2007; Shanks et al., 2003; Willingham et al., 1997), the duration of the RSI does not affect sequence learning as such but the preparation of the response; sequence learning would be relatively equivalent across different RSI values. According to the Awareness hypothesis, the duration of RSI affects the development of explicit knowledge of the target sequence, and therefore the learning of the structural rules underlying the sequence (Cleeremans & Sarrazin, 2007; Destrebecqz & Cleeremans, 2001, 2003; Frensch & Miner, 1994; Huang et al., 2017; Kuhn & Dienes, 2006; Norman et al., 2007; Savalia et al., 2016; Soetens et al., 2004; Willingham et al., 1997). Learning becomes increasingly explicit with increasing RSI duration and remain largely implicit at short RSI duration. Consequently, knowledge requiring conscious elaboration such as structural rules could only be acquired when the RSI is sufficiently long.

In the present study, we tested the predictions of each of these hypotheses by implementing sequences generated by the Fib grammar in the SRT task. The use of these sequences allowed us to quantify learning during processing without having to expose participants to alternative sequences. We manipulated the duration of the RSI in three experiments: 1000 ms in Experiment 1, 250 ms in Experiment 2 and 100 ms in Experiment 3. In all three experiments, RTs for disambiguated points decreased through exposure more than RTs for non-disambiguated points at levels 0, 1 and 2. At level 3, this effect was only present in Experiment 2 where the RSI lasted 250 ms. In all experiments, accuracy decreased systematically for non-disambiguated points. In contrast, accuracy for disambiguated points either increased through exposure or decreased, but to a lesser extent than non-disambiguated points. These results suggest that participants built a hierarchical structure up to the second hierarchical level when the RSI lasted 1000 ms and 100 ms and reached the third hierarchical level when the RSI lasted 250 ms. Taken together with our previous finding that participants also reached the third level with an RSI of 500 ms (Schmid et al., 2023a), it seems that the duration of RSI has a non-linear effect on sequence learning. This non-linear effect of RSI duration on the height of the hierarchical structure cannot be explained by any of the hypotheses put forward.

If the RSI affects sequence learning through stimulus decay in WM, then merging points/constituents should be more difficult as the RSI is longer. Therefore, the height of the hierarchical structure should increase with shortening of the RSI. The Decay hypothesis cannot explain the fact that the hierarchical structure elaborated by participants in Experiment 3 (100 ms RSI) is lower than in Experiment 2 (250 ms RSI) and in Schmid et al. (2023a) (500 ms RSI). Furthermore, exposure was more than twice as long when the RSI lasted 100 ms, this higher exposure should have strengthened the trace of stimuli in WM, and thus promoted hierarchical elaboration.

According to the Preparation hypothesis, RSI duration do not affect sequence learning but the preparation of the response, therefore the height of the hierarchical structure built by the participants (i.e. that reflect sequence learning) should be identical across different RSI duration. The non-linear effect of RSI duration on learning we observed cannot be accounted for by this hypothesis. However, according to this hypothesis, RSI duration affect the preparation of the responses as it modulate the time to prepare for the next trial. This should result in participants being faster overall the longer the RSI and, conversely, the overall slope of RTs should be steeper the shorter the RSI. To test this second prediction, we compared the average RTs and the overall slope of decrease of RTs of the three experiments. We also included in this analysis the results of our previous study where the RSI lasted 500 ms (Schmid et al., 2023a). We only took into account the first 1008 trials in order to have the same number of trials in each experiment (1008 corresponds to the number of trials in Experiment 1 which was the shortest). The results showed that participants were significantly slower when the RSI was 100 ms long compared to all other experiments. There was no difference between the 1000 ms, 500 ms and 250 ms RSIs after the p-values have been corrected for multiple testing. We also found that the decrease in RTs across exposure was greater when the RSI lasted 500 ms and 250 ms compared to when it lasted 1000 ms and 100 ms. There was no difference in slope between the 500 ms and 250 ms RSIs and between the 1000 ms and 100 ms RSIs. These results cannot be explained by the Preparation hypothesis which predicts that the effect of RSI duration on slopes is linear. These results also address a potential confound regarding the way we assessed learning. Indeed, we considered that reaching a hierarchical level results in a larger decrease in RTs for disambiguated points compared to non-disambiguated points of the same level. To the extent that the Preparation hypothesis also predicts an effect on slopes, it could be that the effect of the RSI on the height of the hierarchical structure is in fact due to the preparation of the response by the participants. However, this explanation cannot account for the fact that the overall slopes of the RTs as well as the height of the hierarchical structure built by the participants were non-linearly affected by the duration of the RSI.

Finally, according to the Awareness hypothesis, the length of the RSI would affect the type of knowledge that can be acquired. If hierarchical learning in the Fib grammar involves higher-order structural rules, then the height of the hierarchical structure should increase with the lenghtening of the RSI. Our results can be interpreted in two different ways, depending on the assumption retained. If the Awareness hypothesis is true, i.e. if RSI duration affects the implicit aspect of learning and that structural rules can only be acquired explicitly, since RSI duration affected learning in a non-linear way, our results mean that Fib grammar processing does not require conscious elaboration of structural rules. Note that this is not in contradiction with the hypothesis that the Fib grammar gives rise to hierarchical elaboration. Indeed, we do not claim that the participants have learned the rewriting rules of the Fib grammar in order to access its hierarchical structure, nor that the knowledge they have developed is akin to abstract structural rules. Our hypothesis is that participants build a hierarchical structure from the input by recursively merging points/constituents that span across a deterministic transition. Because the sequence generated by the Fib grammar are aperiodic and self-similar, this mechanism results in a hierarchical structure similar to the natural constituent structure of the Fib grammar. The hierarchical elaboration is thus driven by the particular distributional regularities of Fib and not by the fact that the participants would have acquired the underlying rules of the grammar. If one assumes that hierarchical building in Fib necessarily requires the elaboration structural rules, either participants have acquired these rules implicitly, or RSI duration does not affect the implicit/explicit aspect of learning in the SRT task. However, our results cannot disentangle between these hypothesis since we did not assess participants' awareness of Fib. Note that even if this had been the case, it is not certain that checking participants' consciousness could have provided useful information. Indeed, this evaluation is indirect and is done by adding additional tasks after the learning phase. These tasks therefore measure what remains in memory after processing and not the degree of awareness of the mechanisms involved during encoding.

In summary, none of the hypotheses put forward can explain the non-linear effect of RSI duration on performance. The first question raised by these results is why a non-linear effect of RSI has, to our knowledge, never been reported in the literature. One possible explanation could be that this is due to sampling bias. If, as is often the case, only two RSI durations are compared, then the non-linear effect of RSI is invisible. Our observation would simply come from the fact that we compared more than two RSI durations. While this non-linear effect of the RSI may never have been observed due to sampling bias in RSI duration, this still does not explain its existence.

In the following, we interpret these findings through the lens of recent applications of Shannon's information theory to sequence processing (Pothos, 2010; Radulescu et al., 2019, 2021; Shannon, 1948). According to Shannon information theory (1949), if the amount of information in a signal exceeds the encoding capacity of the receiver, another encoding method should be used to limit the loss of information. Changing the encoding method means compressing the input signal into another format. According to the *Information Premise* (Pothos, 2010), the cognitive system would tends to represents new information with as little uncertainty as possible. To accomplish this, the input is recoded (i.e. compressed) in a way that minimizes the entropy of the system's representational state. According to Radulescu et al. (2019, 2021), acquisition of higher order knowledge result from the tension between the amount of information contained in the input and the cognitive system's limited encoding power in term of memory and processing speed. The encoding power of the cognitive system is defined as the amount of information per unit of time that it is able to process. When the signal exceeds the encoding power, it can be encoded with minimal loss as long as a sufficiently efficient compression method is available. If there is no compression method suitable for the amount of information in the input, the loss of information will increase. Thus, increasing the volume of information per unit of time delivered to the cognitive system compels it to compress the input into a more abstract format as long as the system has a sufficiently powerful compression method at its disposal. If the amount of information delivered to the system exceeds the most efficient compression method, this will result in a loss of information.

In what follows, we adopt this proposal to explain the non-linear effect of RSI duration. We consider that hierarchical elaboration is the manifestation of information compression. Thus, the height of the hierarchical structure elaborated by the participants reflects the degree of compression of the sequence. RSI duration determines the volume of information per unit of time that is delivered to the participants. This hypothesis explain the non-linear effect of RSI duration in the following way: when the RSI lasts 1000 ms, the amount of information to encode per unit of time would not require compressing the sequence beyond the second hierarchical level. As the RSI shortens, the amount of information per unit of time increases and the sequence is further compressed, thus explaining why participants reach the third hierarchical level with RSIs of 500 ms and 250 ms. When the RSI lasts 100 ms, the volume of information is too large to compress the sequence without loss, which explains why participants reached only the second hierarchical level.

4.8 Conclusion

The results of the present study broadly replicate previous observations that participants extract a hierarchical structure when processing Fib grammar sequence in the SRT task (Schmid et al., 2023a; Vender et al., 2019, 2020). This adds to the growing interest in Fibonacci grammar processing (Geambaşu et al., 2016, 2020; Krivochen et al., 2018; Shirley, 2014). Our results also suggest that there is an optimal temporal window for sequence learning in the SRT task. It is possible that these results stem from a tension between the limited encoding power of the cognitive system and the amount of information per unit of time delivered to the system. An open question is whether this non-linear effect of the RSI is specific to the sequence generated by the Fib grammar or whether it can be replicated in other types of sequences. Future work is therefore necessary.

Chapter 5. General Discussion

5.1 Summary of the thesis

The central question of this thesis is whether and how the cognitive system extracts hierarchical structure from a sequentially presented input. Since several underlying structures are possible for a given signal (i.e., several generators with different properties can produce the same signal), the information contained in a signal does not fully determine its structure. This means that the structure extracted by the cognitive system (whether linear or hierarchical) will not depend on the generator of the signal but on the information expressed *in* the signal. The question is therefore to determine which properties of the input are used to extract hierarchical structure. Moreover, in order to know whether the output of the cognitive system reflects the extraction of hierarchical structure, it is not sufficient that the input signal originates from a "hierarchical" generator: one needs to ensure that this output can be obtained from the input only by hierarchical processing.

To explore hierarchical structure extraction, a highly simplified language based on a lexicon of two symbols and a grammar consisting of two rewrite rules was used in the present work: the Fib grammar. This grammar generates binary sequences that are self-similar and aperiodic. The conjunction of these two properties allows us to assess hierarchical learning while controlling for the use of low-level strategies like detecting recurring patterns. The simplicity of this grammar, which nevertheless involves complex formal properties, allows making clear predictions as to how hierarchical learning will manifest in performance. In all the experiments reported in this thesis, strings strings generated by the Fib grammar were presented to the participants through an SRT task. The use of this task allows us to assess learning on-line during encoding, thus avoiding the use of ungrammatical strings and explicit grammaticality judgments that are used in classical AGL tasks. Three specific questions were explored. The first is whether participants process Fib grammar strings as nested structures. The second is the extent to which some of the specific formal properties of the Fib grammar play a role in hierarchical structure extraction. The third concerns the effect of presentation rate on sequence learning in the SRT task. The Fib grammar was used to disentangle three hypotheses about the mechanisms underlying this effect. These questions have been investigated in Chapters 2, 3 and 4 respectively.

In Chapter 2, we tested the hypothesis that participants build hierarchical structure by recursively merging deterministic transitions. Due to self-similarity, transitions between units at level n are identical to transitions between constituents at level n+1 in Fib-generated strings. At each level, transitions are either probabilistic or deterministic. The key feature is that probabilistic transitions at level n are embedded in deterministic transitions at level n+1. It is therefore possible to reduce the number of probabilistic transitions by recursively embedding deterministic transitions. This recursive embedding mechanism makes precise predictions about the pattern of anticipation of points corresponding to a given hierarchical level. Since the number of hierarchical levels is theoretically infinite (of course, for an infinite string only, in our experiment, the maximum number of levels was 12, which is presumably well beyond the processing capacity of the cognitive system), there will always remain higher-order deterministic transitions that are not mastered by the participants, making it possible to determine the maximum height of hierarchical structure extraction that they reached. Results showed that participants' pattern of anticipation could not be accounted for by "flat" statistical learning processes and was consistent with them anticipating upcoming points based on hierarchical assumptions. We also found that participants were sensitive to structure constituency, suggesting that they organized the signal into embedded constituents. The results are compatible with the hypothesis that participants built this structure by recursively merging deterministic transitions.

In chapter 3, we explored whether the isomorphism between surface and derivational properties of the Fib grammar is at the origin of hierarchical structure extraction. According to the k-points hypothesis (Krivochen et al., 2018; Vender et al., 2020), this formal property plays a crucial role in the processing of the Fib grammar whereas according to the recursive merge hypothesis developed in Chapter 2, this factor should, a priori, play no role. We compared the predictions of each of these hypotheses in the processing of the Fib grammar and the Skip grammar. The Skip grammar also belongs to the Lindenmayer formalism: its rewriting rules correspond to two non-consecutive generations of the Fib grammar. As a result, Skip displays similar surface properties (in terms of statistical distribution) while not exhibiting the Fib isomorphism. The results show that in both the Fib and the Skip grammars, participants elaborated a hierarchical structure from the input. This suggests that isomorphism is not a critical factor in hierarchical structure extraction, invalidating the k-points hypothesis. In contrast, the recursive merge hypothesis was found to be compatible with performance on both the Fib and the Skip grammars, with one exception. In the Skip grammar, we found a reverse ambiguity effect at level 4 (i.e., better anticipation for non-disambiguated points than for disambiguated points). We hypothesized that this effect might stem from the difference in the type of self-similarity of the two grammars. Fib displays strong self-similarity (i.e., transitional probabilities are identical at all hierarchical levels) whereas Skip displays weak self-similarity (i.e., transitional probabilities vary with the hierarchical level). As a result, some levels are less complex than others in Skip: the proportion of non-disambiguated points relative to disambiguated points varies among hierarchical levels. Crucially, complexity drops at level 3 of Skip: thus, instead of building the fourth hierarchical level, participants may have resorted to an alternative parsing strategy based on the periodic alternation of the two constituents of level 3. This strategy would be relatively effective in anticipating future points as it would be inaccurate only 5 times per block, and it accounts for the reverse ambiguity effect at level 4.

Finally, in Chapter 4, we explored the impact of presentation rate on hierarchical learning. Three hypotheses have been put forward in the literature to explain the impact of RSI duration in the SRT task : the Decay hypothesis (Frensch & Miner, 1994; Soetens et al., 2004), the Awareness hypothesis (Cleeremans & Sarrazin, 2007; Destrebecqz & Cleeremans, 2001, 2003; Frensch & Miner, 1994; Huang et al., 2017; Kuhn & Dienes, 2006; Norman et al., 2007; Savalia et al., 2016;

Soetens et al., 2004; Verwey & Dronkers, 2019; Verwey & Wright, 2014; Willingham et al., 1997) and the Preparation Hypothesis (Willingham et al., 1997). None of these hypotheses directly focuses on the extraction of hierarchical representations; they address sequence learning in general. To bring new insight to this debate, we took advantage of the fact that the Fib grammar allows us to measure learning in a continuous way without having to resort to the transfer method, which is the basis of most of the studies supporting the aforementioned hypotheses. In three experiments, we manipulated RSI duration and found that it affected performance in a non-linear way, as participants built higher hierarchical structure with an RSI of 250 ms compared to RSIs of 1000 ms and 100 ms. The results suggest the existence of an optimal temporal window for sequence learning in the SRT task, which cannot be accounted for by any of the hypotheses. We hypothesized that this non-linear effect of RSI duration is due to the interaction between the limited encoding power of the cognitive system and the amount of information per unit of time delivered to the system.

The remainder of the Discussion is divided into three sections. In Section 5.2, we assess the evidence supporting hierarchical learning in Fib in light of the results of the experiments conducted in Chapters 2, 3, and 4. We conclude this section with a proposal for an experiment to further investigate hierarchical learning in Fib. In Section 5.3, we discuss our claim that participants segmented Fib into natural constituents and explore whether alternative segmentations could account for the results of the Fib experiments. This section concludes with a proposal for an experiment aiming at further investigating how participants segment a Fib string. In Section 5.4, we investigate whether the model PARSER (Perruchet & Vinter, 1998) which is insensitive to deterministic transitions can segment Fib into natural constituents. Finally, section 5.5 provide concluding remarks for this thesis.

5.2 Do participants extract hierarchical structure in the processing of the Fib grammar ?

The central question of this work is whether the cognitive system extracts a hierarchical structure from string generated by the Fib grammar. To answer this question, one needs an hypothesis about how hierarchical learning should manifest itself in the data. Such hypothesis necessarily pertains to the mechanisms by which a hierarchical structure could be extracted in the Fib grammar. Therefore, the question of whether a hierarchical structure is extracted and the question of how this structure is extracted are intrinsically linked. In the present work, we made the hypothesis that hierarchical structure extraction in Fib would be done by the recusive merge of transition probabilities. In the following, we review the evidence in favour of this alleged processing mechanism in the light of all the chapters. We then justify why the K-point hypothesis (Krivochen et al., 2018; Vender et al., 2020) seems ill-suited to account for hierarchical learning in Fib. We end this section with a suggestion for future research that could provide further evidence for hierarchical processing in the Fib grammar. The rationale for this suggestion is relatively simple: If the non-linear effect of RSI duration reported in Chapter 4 is due to participants building a hierarchical structure, then this non-linear effect of RSI should not be observed if a sequence that does not allow for hierarchical building is used.

5.2.1 Main results in favour of hierarchical learning

To find out whether processing of the Fib grammar results in the creation of hierarchical structure, we conducted several experiments where strings generated by the Fib grammar were implemented in a SRT task. Since these strings are self-similar, transitions between points at level n are identical to transitions between constituents at level n+1. The key feature is that probabilistic transitions at level n are embedded in deterministic transitions at level n+1. It is therefore possible to reduce the number of probabilistic transitions by recursively embedding deterministic transitions. We made the hypothesis that the participants would build a hierarchical structure via the recursive merge of deterministic transitions. The operating principle of the recursive merge hypothesis is relatively simple: the parser would start by merging the points linked by a deterministic transition and would use the resulting constituent to detect new higher-order deterministic transitions. To the extent that some ambiguous points at level n appear at level n+1 in a constituent that follows a deterministic transition, a subset of those ambiguous points can be disambiguated at level n+1. This hypothesis makes precise predictions about the anticipation pattern of points compatible with the elaboration of a hierarchical structure. We tested the predictions of this hypothesis in two ways. If participants build a hierarchical structure by recursively merging deterministic transitions, we should observe: (a) a processing advantage that increases through exposure for points that follow a deterministic transition at level n (i.e. disambiguated points) compared to points that follow a probabilistic transition at the same hierarchical level (i. e. non-disambiguated points), and (b) a better anticipation for disambiguated points appearing at level n+1 in a constituent following a deterministic transition (non-ambiguous structural context) compared to the same disambiguated points occurring at level n+1 in a constituent following a probabilistic transition (ambiguous structural context).

We tested prediction (a) of the recursive merge hypothesis in five SRT experiments using the Fib grammar. The experiments varied along four parameters: the duration of the RSI, the number of blocks, the total number of trials and the number of trials per block. Table 1 displays these parameters for each experiments along with the corresponding number of participants, the chapter in which the experiments are described and the highest hierarchical level reached by the participants. To ensure clarity, we have assigned alphabetical labels to each experiment based on their chronological description, the labels are showin in Table 1.

Table 1

Summary of the SRT experiments using the Fib grammar by Chapter, RSI duration in ms, number of experimental blocks, total number of trials, number of participants and highest hierarchical level reached at the group level.

Chapter	Experiment	Label	RSI duration (ms)	N° of Blocks	N° of trials per Block	Total number of trials	N° of participants	Highest hierarchical level reached
2	Experiment 1	A	500	5	233	1165	159	Level 3
3	Experiment 1	B	1000	5	120	600	53	Level 2
4	Experiment 1	C	1000	7	144	1008	174	Level 2
4	Experiment 2	D	250	5	233	1165	145	Level 3
4	Experiment 3	E	100	7	377	2639	94	Level 2

The processing advantage for disambiguated points over non-disambiguated points was replicated up to level 2 in all experiments, representing in total 625 participants. This advantage also extended to the third hierarchical level in Experiments A and D, representing in total 304 participants. Due to the big sample size, it is unlikely that these effects are due to a type I error. The results of the five experiments using the Fib grammar reported are thus fully in line with prediction (a).

However, as explained in Chapter 2, these results are also potentially compatible with a non-hierarchical processing mechanism sensitive to linear precedence. This explanation relies on the fact that disambiguated points are systematically preceded by a specific sub-sequence that never appears before non-disambiguated points of the same level. In contrast, transitions between sub-sequences of identical length and their following non-disambiguated points are probabilistic. For example, the sub-sequence "0101" always precedes disambiguated points of the second hierarchical level and predicts them with certainty ($p(1|0101) = 1$), whereas the sub-sequence "1101" which always precedes non-disambiguated points of the same level predicts them 38% of the time ($p(1|1101)= .38$). Therefore, the processing advantage of disambiguated points may stem from linear precedence and not hierarchical elaboration. Nevertheless, accounting for the results with linear precedence presents several challenges. First, it would require significant memory resources, as the linear sub-sequences needed to anticipate disambiguated points overlap, necessitating parallel tracking of multiple patterns (see Fig. 5 in Chapter 2 section 2.1). Second, due to the binary nature of the strings, the different sub-sequences can only be distinguished based on their positional order. Therefore, the parser must be able to handle the interference resulting from the similarity between the pattern elements. Finally, the pattern allowing for the anticipation of disambiguated points would need to be held in memory for a relatively long time. This retention time must include the duration of the RSI and the time required to respond to a trial. Table 2 shows the retention times required for linear sub-sequences in function of RSI duration, assuming an average reaction time of 300 ms per trial. In Experiment E, where the RSI lasts 100 ms and participants have reached level 2, the retention time required for anticipating disambiguated points using linear precedence is 1.6

seconds. For Experiments B and C, where the RSI lasts 1000 ms and participants have reached level 2, the sub-sequences must be held in memory for approximately 5.2 seconds. In Experiments A and D, where participants have reached level 3, the sub-sequences must be held in memory for 5.6 and 3.8 seconds, respectively. Thus, a linear precedence parser would need to overcome overlapping patterns, interference from item similarity, and long retention times in working memory to account for the results. The attentional cost imposed by these constraints raises doubts that a simple pattern recognition mechanism would be a feasible approach to explain anticipation of disambiguated points.

Table 2

Retention time (in ms) required to anticipate disambiguated points based on linear precedence by hierarchical level and RSI duration.

RSI duration	Retention time required		
	Level 1	Level 2	Level 3
1000 ms	2600 (ms)	5200 (ms)	9100 (ms) *
500 ms	1600 (ms)	3200 (ms)	5600 (ms)
250 ms	1100 (ms)	2200 (ms)	3850 (ms)
100 ms	800 (ms)	1600 (ms)	2800 (ms) *

Note. The * sign indicate that participants did not reached level 3 with an RSI of 1000 ms (Experiments B and C) and with an RSI of 100 ms (Experiment E).

In regards to prediction (b), Experiment A revealed a processing advantage for points in non-ambiguous structural contexts compared to points in ambiguous structural contexts at levels 1 and 3. This is in line with the recursive merge hypothesis. Since ambiguous and non-ambiguous structural contexts contrast different instances of disambiguated points only (i.e., no non-disambiguated points were included in this analysis), points in both structural contexts were preceded by identical sub-sequences (with a transition probability equal to 1). Therefore, effects of structural contexts cannot be explained by a strategy consisting of anticipating disambiguated points based on linear precedence.

5.2.2 Alternative account of hierarchical learning in Fib

If participants are extracting a hierarchical structure from Fib string, they may not necessarily do so by the recursive merge of deterministic transition. The k-points hypothesis (Vender et al. 2020) proposes an alternative explanation as to how hierarchical structure is extracted in Fib. The k-points hypothesis is based on the formal approach developed by Krivochen (Krivochen et al., 2018) who hypothesized that the parser would take advantage of the isomorphism between the surface properties of the Fib grammar and its structural properties in order to extract a hierarchical structure. According to this hypothesis, the identification of specific points in the grammar, called k-points, would enable the parser to build the local hierarchical structure of the grammar because of their specific structural status (see Chapter 3 section 3.1 for detail). However, many of our observations cannot be explained by this hypothesis. First, if k-points' formal status is the basis for building the hierarchical structure, there should be a processing advantage for all k-points as they should be identified in the same way. However, our results are in contradiction with this prediction. In our notation, k-points are the non-disambiguated points of level 0, and therefore, level 2 contrasts different instances of k-points. Consequently, the better anticipation of disambiguated points compared to non-disambiguated points at level 2 that we replicated in Experiments A, B, C, D and E cannot be explained by this hypothesis. Second, if hierarchical learning in Fib is based solely on the identification of k-points, the processing advantage should be restricted to k-points. Again, our result show that this is not the case. In particular, the disambiguation effect found at level 3 in Experiment A and replicated in Experiment D, which pertains to differences between 0s, cannot be accounted for by this hypothesis since k-points are, by definition, 1s. In summary, results support our hypothesis that when processing the string generated by the Fib grammar, participants build a hierarchical structure through the recursive merge of deterministic transitions rather than by the identification of K-points.

5.2.3 Is the non-linear effect of RSI duration a marker of hierarchical learning?

In line with Isbilen & Christiansen, (2022) the summary in Table 1 suggests that length of exposure has no effect on learning. This observation aligns with our statement in Chapter 4, where we observed that increasing exposure did not seem to improve hierarchical learning. Our justification was based on the fact that, despite a more than twofold increase in exposure in Experiment E as compared to Experiment D, participants only reached the second hierarchical level in Experiment E, while they reached the third hierarchical level in Experiment D. However, this comparison was improper as RSI duration differed between the experiments. Nevertheless, a comparison between Experiments B and C eliminates this confounding factor and supports the hypothesis that length of exposure does not improve hierarchical learning: even though the exposure was 40% longer in Experiment C than in Experiment B, participants in both experiments only reached the second hierarchical level. Rather, RSI duration seems to be the factor that determines the height of the hierarchical structure built by the participants.

The non-linear effect of RSI duration on performance may provide interesting insight on the mechanisms involved in the processing of the Fib Grammar. In particular, it suggests the existence of an optimal temporal window for learning. In Chapter 4, we saw that the height of the hierarchical structure elaborated by the participants was higher when the RSI lasted 500 ms and 250 ms and decreased with an RSI of 1000 ms and 100 ms. This U-shape effect has, to our knowledge, never been reported, possibly because of a sampling bias in RSI duration as most studies compared only two RSI durations. However, the mechanism underlying this effect remains to be unveiled as none of the hypotheses proposed in the literature could account for it.

To account for it, we have considered hierarchical elaboration as an information compression mechanism. According to Shannon information theory (1948), if the amount of information in a signal exceeds the encoding capacity of the receiver, another encoding method should be used to limit the loss of information. Changing the encoding method means compressing the input signal into another format. According to the *Information Premise* (Pothos, 2010), the cognitive system would always try to represent new information with as little uncertainty as possible, that is, minimizing the entropy of its representational state. The tension between the limited processing capacity of the cognitive system and the amount of information contained in a signal would be the driving force that leads the cognitive system to develop higher-order knowledge (Radulescu et al. 2019, 2021). The idea is as follows: the encoding power of the cognitive system is limited by its processing speed and memory storage capacity. The encoding power can be defined as the amount of information per unit of time that the system is able to process. When the input exceeds the encoding power, it can be encoded with minimal loss as long as a sufficiently efficient compression method is available. If there is no compression method suitable for the amount of information in the input, the loss of information will increase. In this framework, the duration of the RSI determines the amount of information per unit of time that is delivered to the participants. If we assume that the height of the hierarchical structure elaborated reflects the degree of compression of the input, the U-shape pattern is explained in the following way: when the RSI lasts 1000 ms, the pressure on the system would not be high enough for it to elaborate the third hierarchical level. As the RSI shortens, the amount of information per unit of time increases and the input is further compressed, thus explaining why participants reach the third hierarchical level with RSIs of 500 ms and 250 ms. With an RSI of 100ms, the amount of information per unit of time would be too high without involving information loss: participants would not have a sufficiently efficient compression method to encode so much information, which explains the decrease in the height of the hierarchical structure.

5.2.3.1 Idea for future work: exploring the effect of RSI duration in other types of sequences

This hypothesis makes an interesting prediction regarding the question of whether the processing of the Fib grammar actually results in the elaboration of hierarchical structure. Indeed, this hypothesis assumes that the presentation rate affects hierarchical elaboration, which consists in input compression. Crucially, a non-hierarchical processing mechanism is incompatible with the notion of compression: if disambiguated points are anticipated on the basis of linear precedence in a purely item-based manner, they cannot, by definition, be compressed. Thus, if the compression of the input is the cause of the non-linear effect of RSI duration, a sequence in which there would be no hierarchical structure to build upon should not exhibit this effect. Testing that prediction could be done with an experiment in which target sequences give rise to "flat" statistical learning without hierarchical learning. One possible approach would be to use a sequence that follows a Second Order Conditional (SOC) structure (Reed & Johnson, 1994).

5.3 Does the parser identify Fib-natural constituents?

In this section, we ask whether our results can be explained without assuming that the participants built the natural constituents of Fib. According to the recursive merge hypothesis, participants build a hierarchical structure by recursively merging points/constituents that span across a deterministic transition. This process results in a nested structure consisting of the natural constituents of the Fib grammar. The reason why participants would have developed the natural constituents and not others is a consequence of the interaction between the alleged operating principle (i.e., recursive merge) and the statistical properties of Fib grammar strings. Since at each hierarchical level there is exactly one deterministic transition, recursive merge of deterministic transitions can only result in the building of the natural Fib constituents. As we have seen in Chapter 2, the patterns of anticipation of points at the group level are replicated at the individual level. We interpreted this finding as suggesting that participants had all formed a hierarchical structure made up of natural Fib constituents. However, this rationale holds only if the segmentation into natural constituents is the only one that predicts the observed anticipation patterns. If there are other ways of segmenting a Fib string that yield identical predictions, the precise constituents built could vary from one participant to another despite the overall similarity of the anticipation patterns among participants. In the following, we explore whether the segmentation into natural Fib constituents is the only one that can account for the patterns of anticipation observed in experiments A, B, C, D and E.

5.3.1 Classification of segmentations in Fib

Given the high number of possible segmentations, we had to find a criterion to restrict our analysis. For this purpose, we took advantage of the self-similarity of the Fib grammar. Because of the self-similarity, any Fib generation can be segmented into a pair of n-grams of length equal to two consecutive numbers of the Fibonacci sequence (as a reminder, each number of the Fibonacci sequence is the sum of the two previous ones: 0-1-1-2-3-5-8, etc.). Segmenting a Fib string in this way minimizes the number of different n-grams needed to ensure there are no remainders: only two n-grams are needed for optimal tessellation. If n-grams are used whose length is not a number of the Fibonacci sequence (for instance a 4-gram), at least three different n-grams are required for the full mapping of the string. Let's consider the eighth generation of the Fib grammar shown in (i). The first possible segmentation involves using the natural constituents of Fib, as illustrated in (ii) using the 2-gram [01] and the 3-gram [101] that correspond to level 2. In this case, both constituents perfectly match the sequence, leaving no "remainder". However, we can also segment (i) without leaving a remainder with another pair of 2-gram and 3-gram. As an example, (iii) shows the seventh generation segmented using the 2-gram [10] and the 3-gram [101], which perfectly maps the sequence.

(i) 101011010110110101101

(ii) [101][01][101][01][101][101][01][101]

(iii) [10][101][10][101][101][10][101][101]

The question is therefore whether there are some alternative segmentations that are compatible with our results. Since the number of possible segmentations increases exponentially with the size of the constituents considered, we will restrict our analysis to the set of segmentations consisting of a 1-gram and a 2-gram, which is equivalent to the first hierarchical level, and to the set of segmentations consisting of a 2-gram and a 3-gram, which is equivalent to the second hierarchical level.

5.3.1.1 Segmentations with a 1-gram and a 2-gram

There are only two options for segmenting a Fib string with a 1-gram and a 2-gram. The first is to use the natural Fib constituents [1] and [01], as shown in (iv). The second is to use the 1-gram [1] and 2-gram [10], as shown in (v). We have highlighted in green the points that are disambiguated by each segmentation and in red the points that are not. Upon examination of (v), it appears that the 1-gram [1] is consistently followed by the 2-gram [10]. Therefore, the points that constitute [10] when preceded by [1] are disambiguated whereas points that constitute [10] when preceded by [10] are non-disambiguated. This segmentation predicts the opposite ambiguity effect at level 0 than the segmentation into natural constituents (i.e., better anticipation for non-disambiguated points compared to disambiguated points at level 0), which is in complete contradiction with our results. Additionally, segmentation (v) makes it impossible for the parser to determine the identity of the point following the 1-gram [1] *a priori* because both n-grams of this segmentation start with a "1" (e.g. [1] and [10]). When a "1" is perceived, how could the parser ascertain whether this "1" marks the beginning of the 2-gram [10] or the 1-gram [1]? Identification of the border between n-grams could thus be done only *a posteriori*, when a "0" is perceived. This makes it impossible to use segmentation (v) regularity “[1] is always followed by [10]” to improve anticipation.

(iv) [1][01][01][1][01][01][1][01][1][01][01][1][01]

(v) [10][10][1][10][10][1][10][1][10][10][1][10][1]

5.3.1.2 Segmentation with a 2-gram and a 3-gram

There are four ways of segmenting a Fib string with pairs of 2-grams and 3-grams. The segmentation into natural constituents of Fib [01] and [101] is shown in (vi). In (vii), we have segmented with the 2-gram [10] and the 3-gram [101]. In (viii), we have segmented with the 2-gram [01] and with the 3-gram [011]. The segmentations (vii) and (viii) both predict the opposite ambiguity effect at the first hierarchical level than the segmentation into natural constituent (i.e., better anticipation for non-disambiguated points compared to disambiguated points at level 1), which contrasts with our finding. Moreover, like (v), the n-grams of segmentations (vii) and (viii) start in the same way: in (vii) the n-grams [10] and [101] both start with "10" and in (viii), n-grams [01] and [011] both start with "01". Therefore, in these segmentations, the parser can only determine the beginning of an n-gram *a posteriori* when processing the next n-gram, making it impossible to use the regularity of segmentation (vii) "[10] is always followed by [101]" or segmentation (viii) "[01] is always followed by [011]" to improve anticipation.

(vi) [101][01][101][01][101][101][01][101]

(vii) [10][101][10][101][101][10][101][101]

(viii) 1][01][011][01][011][011][01][011][01

(ix) 10][10][110][10][110][110][10][110][1

Segmentation using the 2-gram [10] and the 3-gram [110] is presented in (ix). Unlike (vii) and (viii), (ix) yields predictions that align with our findings. This segmentation precisely disambiguates the same set of ambiguous points as (vi). Additionally, the 2-gram [10] and the 3-gram [110] differ enough to enable the parser to discern their location before encountering an ambiguous point. However, the creation of these particular n-grams in the first place is somewhat challenging to explain. The recursive merge hypothesis suggests that participants must first segment the string as in (iv) to access (vi). As (v) cannot lead to (ix), the only plausible explanation is that participants directly built the second hierarchical level. The mechanism underlying this possibility would need to be explained.

In summary, we see that if we restrict ourselves to the second hierarchical level, only two alternative segmentations are compatible with our results: the segmentation into natural Fib constituents (as in (vi)) and segmentation (ix). Of course, the number of compatible segmentations increases as the size of the n-grams considered grows, however, in order to be compatible with our results, upper levels segmentations must adopt segmentation (iv) or segmentation (ix) at level 2. Our results do not provide a definitive answer on which segmentation ((iv) or (ix)) was adopted by the participants. However, although segmentation (ix) fits with the results, it remains to be explained how this segmentation would be attained in the first place. In contrast, the segmentation into natural constituents of Fib is accounted for by a single operating principle: the recursive merge of deterministic transition.

5.3.2 Idea for future work: Further exploration of the constituent structure

To explore further whether the different segmentations mentioned above facilitate or hinder learning, a segmentation scheme could be highlighted by manipulating perceptual features of the stimuli or the location at which they appear. This could be easily achieved since in the experiments conducted in this work, only one variable indexed hierarchical structure (i.e., the order of presentation of the stimuli), leaving a wide range of parameters that could be manipulated. Such manipulation would also permit to highlight several hierarchical levels simultaneously. This would allow us to explore from a different angle the representation elaborated by the cognitive system. i.e., to see if the participants consider all levels of the structure simultaneously. The addition of perceptual cues could therefore open up interesting perspectives for future studies.

5.4 Does the parser identify Fib-natural constituents by detecting deterministic transitions?

The recursive merge hypothesis posits that hierarchical structure extraction in Fib involves recursively merging deterministic transitions. In a Fib string, the boundaries between constituents at each level are denoted by deterministic transitions. By having only one deterministic transition per hierarchical level, the recursive merge method always segments Fib into natural constituents. In order to detect deterministic transitions, it is necessary to assume that the cognitive system computes transition probabilities between the elements of the sequence. However, there is no consensus on whether the cognitive system actually computes transition probabilities when segmenting a signal.

The question of how constituents are extracted from a linear string is actually the same question that underlies the wide debate about how words are extracted from a continuous stream (see Perruchet, 2019 for a review). Two main approaches have been adopted, that assume substantially different mechanisms responsible for detecting the units that make up a signal. The first assumes that the initial stage of signal processing is based on the calculation of pairwise statistics between salient perceptual elements such as transition probabilities (Adini et al., 2015; Meyniel et al., 2016; Saffran et al., 1996). In a seminal study, Saffran et al. (1996) exposed 8-month-old infants to a continuous stream of syllables, which were arranged in triplets that formed pseudo-words. Importantly, the only cue to the boundaries between triplets was the statistical distribution of syllables: transition probabilities between syllables of a word are higher than those between syllables at word boundaries. Following exposure to the speech stream, the infants were tested on their ability to discriminate between triplets that did or did not contain a boundary. The results showed that the infants were able to differentiate between the two types of triplets, suggesting that infants were computing transition probabilities to detect word boundaries. Our recursive merge hypothesis is in line with this assumption that constituent boundaries are identified on the basis of transitional probabilities between units. Our hypothesis further assumes that the cognitive system computes transitional probabilities recursively at higher levels, as constituent structure is being built. The second approach assumes that a chunking mechanism of the input is responsible for the extraction of the units, without calculation of transitional probabilities. This approach was initially proposed and modelled by Perruchet and Vinter (1998), and alternative implementations of the same idea have been proposed subsequently (French et al., 2011; Goldwater et al., 2009; McCauley & Christiansen, 2014; Robinet et al., 2011; Thiessen & Erickson, 2013). The idea is that the system aggregates elementary elements simultaneously present in the focus of attention. This process leads to the creation of chunks that either map or do not map the actual units of the signal. Chunks that correctly map the units will occur more often and therefore be reinforced as processing units. The chunking approach does not state that the statistical distribution of the input does not play a role, but that one does not have to assume that the system computes transitional probabilities to segment the signal.

In the following section, we explore whether it is possible to detect boundaries between natural constituents without relying on deterministic transitions. If a model that does not compute transition probabilities systematically segments Fib into natural constituents, it would show that along deterministic transitions, there is another property in the Fib strings that can be used to reliably access constituent structure. To explore this question, we conducted a series of simulations using the PARSER model developed by Perruchet and Vinter (1998), using as input Fib grammar-generated strings.

5.4.1 Overview of PARSER

The purpose of the PARSER model was to showcase that stream segmentation could be achieved without calculating transition probabilities. The model accomplishes this by randomly grouping units together from the beginning of the learning process. The model operates cyclically according to the steps in Fig. 1, and requires primitives (0 and 1 in our case) as the minimal element of a sequence to function. At each cycle, the span of the focus of attention randomly changes, ranging from 1 to 3 slots. Initially, only primitives are stored in memory, but as the model encounters pairs of units in its input, it creates new units by chunking them together. Once a chunk is formed, it is treated as a single unit within the focus of attention. Therefore, if the same chunk appears again, it occupies only one slot in the focus of attention. This process leads to the creation of chunks that either map or do not map the actual units of the signal. At each cycle, the activation value of all the chunks that were not in the focus of attention declines whereas the activation value of the chunks present in the focus of attention increases. When multiple chunks in memory correspond to the input (i.e., when the signal fits chunks of different sizes), only the chunk with the best fit see its activation value increases. As a result, chunks that correctly map the words in the input will have their activation value increased because they occur more often while chunks that only partially map the input will see their activation value progressively reduced and tend to zero. This model tends to select the segmentation of an input that requires the smallest number of different chunks, as only the best-fitting (longest) chunks see their activation value increase, while it decreases for the least-fitting chunks.

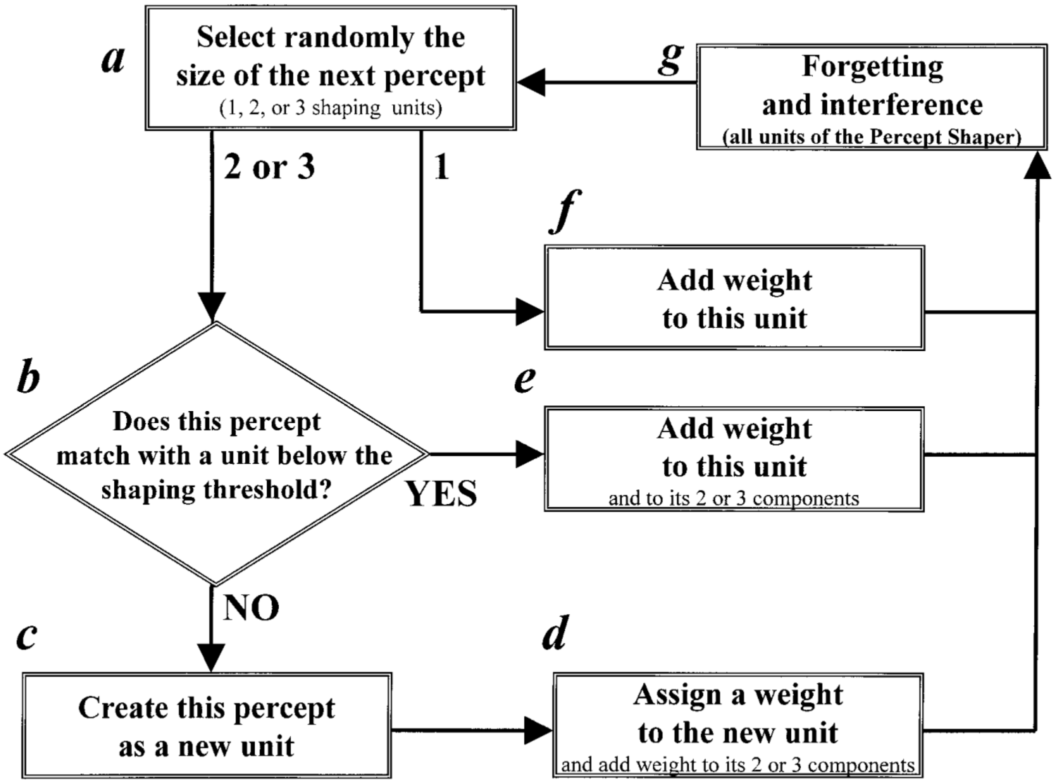


Fig 1. Flow chart of the PARSER model, from Perruchet & Vinter, (1998).

5.4.2 Simulations of Fib string chunking by PARSER

In order to find out whether PARSER segments Fib strings into natural constituents in a systematic way, we conducted a series of simulations where strings of the Fib grammar were presented to the model. The simulations reported below were carried out with the PARSER implementation in the U-learn software (Perruchet et al., 2014). There are several important aspects of the model's functioning that should be highlighted. First, the model's parameters, including decay rate, interference rate, and the minimum and maximum percept sizes, are set by the experimenter. Default parameter values are given in Perruchet and Vinter (1998), and we used these values in our simulations. Second, the results of a simulation are determined by the seed that sets the vector of subsequent sizes of attentional focus. Thus, comparing simulations with the same parameters tells us something about the probability of a parse to occur. Third, the model's implementation does not grant access to the memory's content cycle by cycle, but only at the conclusion of each run. To gauge the memory content's evolution, we employed three input strings of varying lengths: (i) the 11th generation of the Fib grammar with a length of 144 points, (ii) the concatenation of the 11th generation 7 times (1008 points), and (iii) the 20th generation of the Fib grammar (10946 points). Fourth, the software options for exploring the memory content across multiple runs is limited and this had consequences on the way we carried out our investigation. The model's implementation allows tracking up to 10 lists of chunks across multiple runs. A list can contain multiple chunks, however, when it this the case, the model returns only the frequency of the lists (i.e., the number of occurrences where at least one member of the list was present in memory at the end of learning) and not the frequency of the chunks that compose it. Due to the vast number of possible n-grams in Fib (which exceeds 10 by far), the combinations of these n-grams result in a huge number of potential segmentations. As the number of n-grams that can individually be monitored across multiple runs is restricted to 10 (i.e., by creating 10 lists, each comprising a single n-gram), it is clearly impossible to test every possible segmentation. This led us to conduct three analyses where we explored a distinct aspect of the behavior of PARSER. Each aspect explored represents a necessary condition that PARSER must fulfil in order to conclude that the model segments Fib into natural constituents in a systematic way. We explain these analyses below.

The first analysis aimed at establishing whether the model chunks n-grams whose length is a number of the Fibonacci sequence. Indeed, we know that the length of the natural constituents of the Fib grammar correspond to a number of the Fibonacci sequence. Thus, if PARSER segments Fib into natural constituents, n-grams that are not the length of a number of the Fibonacci sequence should not be created by the model. The second analysis aimed at testing whether n-grams of identical length are present simultaneously in the model's memory. The reason behind this analysis is that all natural Fib constituents have a different length, i.e., there is never two n-grams of the same length (with the exception of the primitive 0 and 1). In PARSER, there is no constraint on the number of n-grams of a particular length: it is therefore possible to have multiple n-grams of the same length. Therefore, if PARSER segment Fib into natural constituents, two n-grams of identical lenght should not coexist in the model's memory at the end of the learning process. Finally, in the third analysis, we explored the identity of the Fib length n-grams chunked by the model. If PARSER segments Fib into its natural constituents, then Fib length n-gram that are not natural constituents of the grammar should never be chunked.

In the following section, we report and discuss the results of each analysis. In each analysis, the model was run 1000 times for each input length, resulting in 3000 simulations per analysis. In total, we ran 9000 simulations. In each analysis, the Fib grammar alphabet 0 and 1 were utilized as primitives.

5.4.2.1 Is PARSER chunking n-grams of Fib length?

Fig. 2 displays the frequency of chunks as a function of their length (ranging from 1 to 10), irrespective of their identity. For instance, the blue bar in the 2-gram category indicates that a 2-gram was detected in 696 out of 1000 runs, regardless of whether it was 01, 10, or 11. Each category contains all possible n-grams in Fib for a given length. The list of n-grams in each category is given in appendix S4.

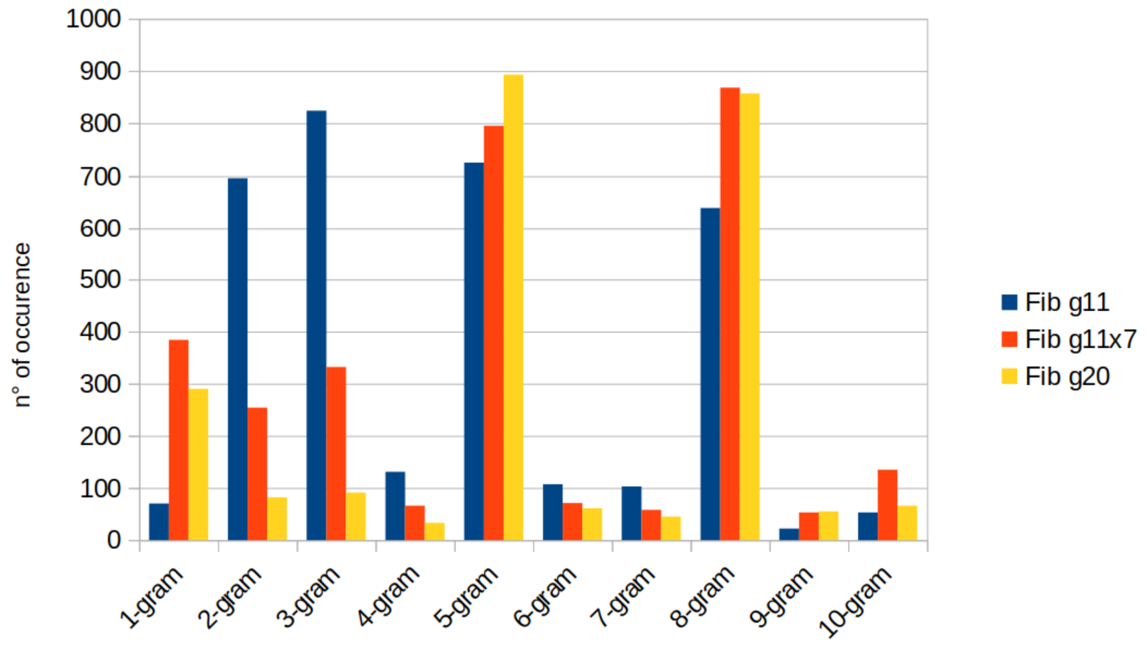


Fig 2. *Frequency of n-grams per size present at the end of the training by input length: generation 11 of the Fibonacci grammar (Fib g11), generation 11 of the Fibonacci grammar concatenated seven times (Fib g11x7), and generation 20 of the Fibonacci grammar (Fib g20).*

Two conclusions can be drawn from this first analysis. First, the frequency of short n-grams diminishes with exposure: 2-grams and 3-grams are more frequent when the model is presented with short strings (Fib g11) compared to medium length strings (Fib g11x7), and even less frequent in long strings (Fib g20). This suggests that the model chunks 2-grams and 3-grams in the initial stages of learning, but they disappear later due to decay. Second, and more importantly, the model rarely chunks n-grams that are not the length of a number of the Fibonacci sequence: n-grams of Fib length (i.e., 1, 2, 3, 5, and 8-grams) are more likely to be chunked compared to n-grams that are not the length of Fib (i.e., 4, 6, 7, 9, and 10-grams). This effect can be explained by the fact that the model favors the shortest lexicon (i.e., the partitioning of a corpus with the smallest number of different words). As seen in section 5.3, optimal segmentation of Fib strings (i.e., segmentations where there are no remainders) involved only two Fib-length n-grams. If a chunk that is not the length of a number of the Fibonacci sequence is created (e.g., a 4-gram), it will inevitably align less well with the input than a Fib-length n-gram and will therefore decay as soon as a Fib-length n-gram is created. The results of this first analysis show that n-grams that are not Fib-length are almost never chunked by PARSER.

5.4.2.2 Do chunks of identical length co-exist in PARSER's memory?

Fig. 3. shows the frequency of simulations in which the model retained two chunks of the same length at the end of the learning. As no instance of three or more identical chunks in memory occurred across the 3000 simulations, the frequency of such occurrences could not be depicted.

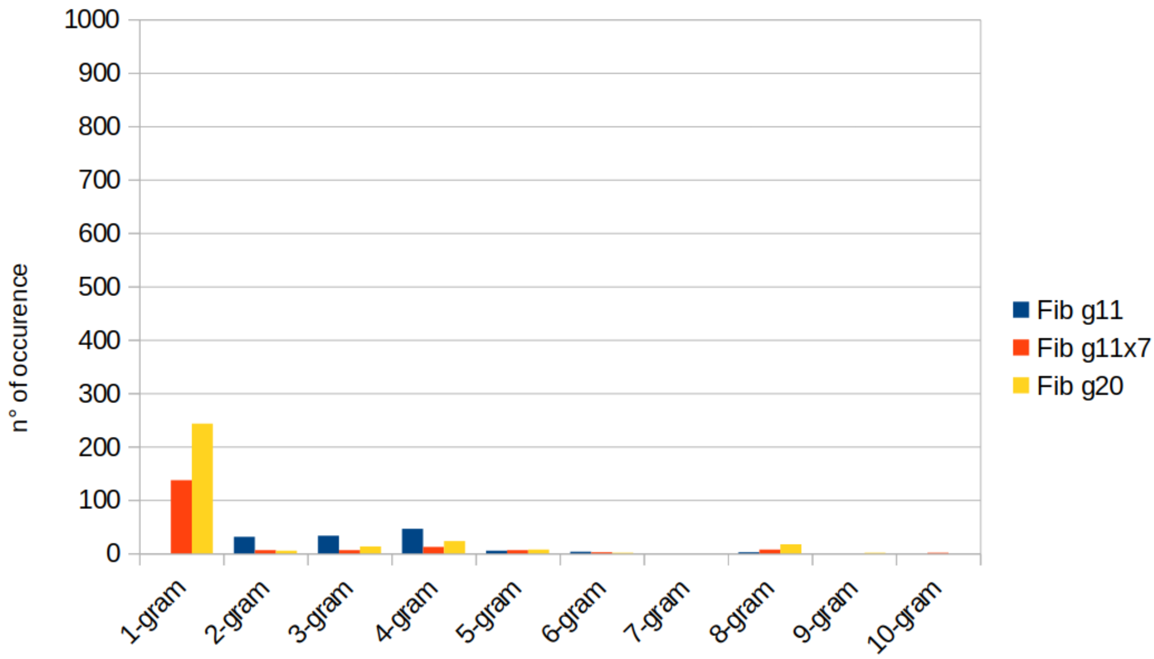


Fig 3. *Absolute frequency of run in which two n-grams of the same length were present at the end of the learning, by input length: generation 11 of the Fibonacci grammar (Fib g11), generation 11 of the Fibonacci grammar concatenated seven times (Fib g11x7), and generation 20 of the Fibonacci grammar (Fib g20).*

The results indicate that the model almost never stores two chunks of the same size simultaneously in memory. This observation can be attributed to the fact that there are optimal ways of tessellating a Fib string (i.e., segmenting without leaving any remainder). The optimal segmentation of a Fib string involves Fib-length n-grams of different lengths. Thus, once the model settles on a particular set of n-grams, these n-grams will be reused over and over again because they tessellate the string in an optimal way, leaving n-grams of identical length that are not in that set virtually no chance of appearing. Note that bigger n-grams can still appear because they provide a better fit of the input.

The results of this second analysis show that PARSER almost never segments two n-grams of the same length at the same time.

5.4.2.3 Is PARSER chunking Fib-natural constituents ?

Fig. 4. shows the frequency of occurrence of all 3-grams and 5-grams possible in the strings generated by the Fib grammar.

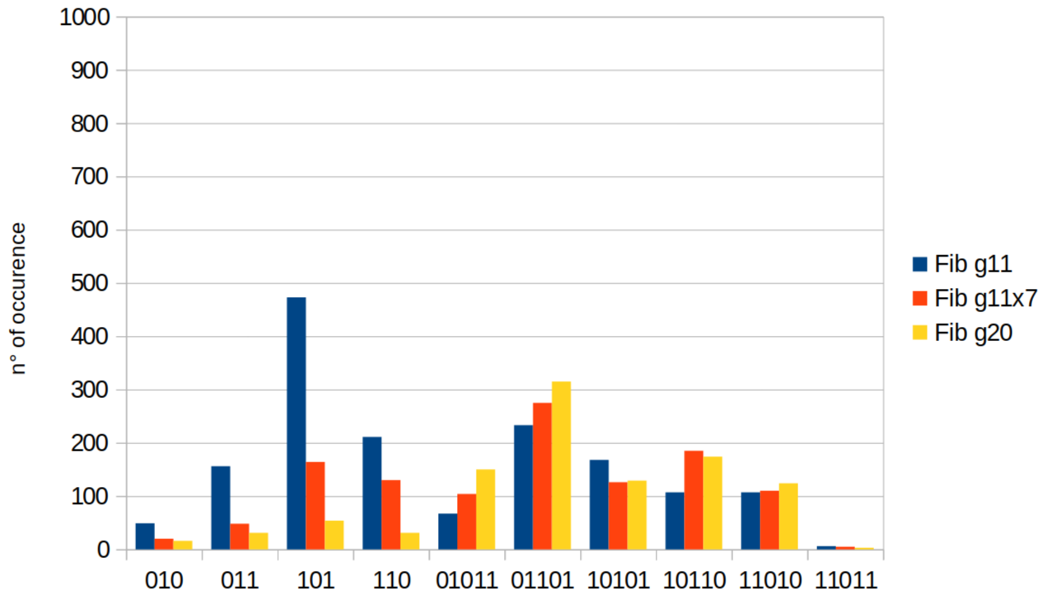


Fig 4. *Frequency of occurrence of all 3-grams and 5-grams in memory at the end of the learning by input length : generation 11 of the Fibonacci grammar (Fib g11), generation 11 of the Fibonacci grammar repeated seven times (Fib g11x7), and generation 20 of the Fibonacci grammar (Fib g20)*

The first observation to note is that the 3-gram [010] and the 5-gram [11011] are chunked significantly less often than the other 3-grams and 5-grams. This is explained by the fact that these n-grams are not sufficient to tessellate a Fib string in an optimal way; they will thus always provide a poorer fit than the other 3-grams and 5-grams, and are therefore less likely to be chunked by PARSER. The second point to note is that even though PARSER chunks the 3-gram [101] and the 5-gram [01101], other 3-grams and 5-grams are chunked in the majority of the simulations. In total, the 3-gram [101] appears in 23.6 % of the simulations (49.1% in Fib g11, 16.4% in Fib g11x7 and 5.4% in Fib g20) and the 5-gram [01101] appears in 27.4% of the simulations (23.4% in Fib g11, 27.5% in Fib g11x7 and 31.5% in Fib g20). An important point to note is that these values overestimate the number of times PARSER actually chunked the natural constituents [101] and [01101]. This is because not all 3-grams [101] and 5-grams [01101] are natural constituents. For instance, the segmentations (vi) and (vii) in section 5.3.1.2 show that the 3-gram [101] is a natural constituent in (vi) but not in (vii) (we have reproduced these segmentations below for ease of reading). Therefore, in the 23.6% of simulations where PARSER chunked the 3-gram [101], only a (unknown) subset corresponds to the natural constituents [101]. The same demonstration can be done for the 5-gram [01101] (and in fact for all natural constituents of a length greater than 2). This also explains why the 3-gram [101] and the 5-gram [01101] are chunked more often, as they may or may not appear as natural constituents, they are overall more frequent than other 3-grams and 5-grams. The results of this third analysis show that in most simulations, PARSER chunks n-grams that are not natural constituents of the Fib grammar.

(vi) [101][01][101][01][101][101][01][101]

(vii) [10][101][10][101][101][10][101][101]

In summary, although the first two analyses showed that PARSER almost exclusively chunk Fib-length n-grams and that two n-grams of identical length almost never coexist in the model memory, the results of the third analysis showed that, although PARSER occasionally identifies natural constituents of the Fib grammar, it does not do so in a systematic way. This suggests that the Fib strings do not possess properties allowing PARSER to reliably detect their constituent structure. It thus seems that in order to systematically segment Fib into its natural constituents, it is necessary to rely on the computation of deterministic transitions.

5.4.3 Can PARSER account for participants' anticipation patterns in Fib?

The fact that PARSER does not systematically segment Fib-generated strings in natural constituents does not mean that this model cannot account for our experimental results. As we saw in section 5.3, there are (few) alternative segmentations that predict a pattern of anticipation of points identical to that predicted by the segmentation into natural constituents. Hence, it is possible that in the simulations where PARSER did not segment the Fib strings into natural constituents, it still performed a segmentation that aligns with our experimental results.

To the extent that some of the segmentations that align with the results share a common n-gram with segmentations that does not, it is difficult to estimate the number of times PARSER performed a segmentation compatible with the results. However, we can easily approximate how often PARSER produced a segmentation that predicts a pattern of anticipation of points that does *not* align with the results. In Fig. 4, all the simulations where PARSER has chunked the n-grams [011], [10101], [10110] or [01011] necessarily falls into a segmentation that predict an anticipation pattern opposite to the one we have observed experimentally. We see that in 44% of the simulations, at least one of these n-grams appears. This percentage underestimates the number of times PARSER actually segmented Fib in a way that is inconsistent with the results because it does not take into account the n-grams that may or may not appear in these segmentations (i.e., n-grams [101], [110], [01101], and [11010]). Nevertheless, it shows that in at least 44% of the cases, PARSER segmented Fib in a way that is inconsistent with the experimental results. In section 2.4 of Chapter 2, we saw that the pattern of anticipation at the group level is replicated at the individual level. Thus, this proportion of 44% seems too high to conclude that PARSER accounts for the anticipation pattern of the participants.

5.5 Conclusion

The different experiments reported in this work strongly suggest that participants extract a hierarchical structure during processing of strings generated by the Fib Grammar. The recursive merge hypothesis provides a satisfactory explanation of our results. Its operation is simple: participants recursively merge points that span across a deterministic transition, and use the output of this process, i.e., the constituents created by this merging process, to detect new higher-order deterministic transitions. The result of this process is a representation of embedded constituents, or a ‘chunk of chunks’ representation. The fact that this mechanism results in a constituent structure identical to the natural constituent structure of Fib is due to the Fib-specific self-similarity which makes the transitional probabilities perfectly scale-free. Thus, the surface properties, i.e., the transitional probabilities, lead the parser to a structure that is identical to the natural structure of Fib.

It is important to stress that the scope of the recursive merge hypothesis is relatively limited. While this hypothesis offers a satisfactory explanation for our findings, it remains to be determined whether its applicability can be expanded beyond fractal binary sequences generated by L-systems. More precisely, the question at hand pertains to whether the recursive calculation of transitional probabilities is contingent upon specific characteristics present in the input, or if it is a multipurpose statistical learning mechanism commonly used by the cognitive system. Future studies testing the recursive merge hypothesis in other types of sequences are therefore needed to answer this question.

A related question is to what extent hierarchical structure extraction is underpinned by a domain-general ability. More specifically, is hierarchical processing in Fib and in other cognitive domains underpinned, at least in part, by a common processing module or are they distinct mechanisms that nonetheless share a set of domain-general principles ? Examining the correlation between participants' performance in Fib and other types of hierarchical processing, such as those involved in natural language processing, could shed light on this inquiry. This perspective presents an intriguing direction for future studies.

In conclusion, although many questions remain unanswered at the end of this work, it nevertheless remains that the Fib grammar and L-systems in general offers an interesting potential for the study of hierarchical processing and the underlying mechanisms involved.

6. References

Adini, Y., Bonneh, Y. S., Komm, S., Deutsch, L., & Israeli, D. (2015). The time course and characteristics of procedural learning in schizophrenia patients and healthy individuals. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00475

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, *52*(1), 68-81. https://doi.org/10.3758/s13428-019-01205-5

Bahlmann, J., & Friederici, A. D. (2006). *FMRI investigation of the processing of simple linear and embedded hierarchical structures : An artificial grammar task*. 126. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_733942

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Lme4 : Linear mixed-effects models using Eigen and S4. R package version 1.1-7.*

Chomsky, N. (1957). Logical structures in language. *American Documentation (pre-1986)*, *8*(4), 284.

Chomsky, N. (1995). The minimalist program MIT Press. *Cambridge, MA*.

Chomsky, N., & Lightfoot, D. W. (2002). *Syntactic Structures*. Walter de Gruyter.

Christiansen, M. H. (2019). Implicit Statistical Learning : A Tale of Two Literatures. *Topics in Cognitive Science*, *11*(3), 468-481. https://doi.org/10.1111/tops.12332

Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness : A graded, dynamic perspective. *Implicit learning and consciousness*, *2002*, 1-40.

Cleeremans, A., & Sarrazin, J.-C. (2007). Time, action, and consciousness. *Human Movement Science*, *26*(2), 180-202. https://doi.org/10.1016/j.humov.2007.01.009

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs : A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*(9), 751-766. https://doi.org/10.1016/j.tics.2022.06.010

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The Neural Representation of Sequences : From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, *88*(1), 2-19. https://doi.org/10.1016/j.neuron.2015.09.019

Destrebecqz, A., & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review*, *8*(2), 343-350. https://doi.org/10.3758/BF03196171

Destrebecqz, A., & Cleeremans, A. (2003). Temporal effects in sequence learning. In L. Jiménez (Éd.), *Advances in Consciousness Research* (Vol. 48, p. 181-213). John Benjamins Publishing Company. https://doi.org/10.1075/aicr.48.11des

de Vries, M. H., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning : The learnability of embedded hierarchical structures. *Cognition*, *107*(2), 763-774. https://doi.org/10.1016/j.cognition.2007.09.002

de Vries, M. H., Petersson, K. M., Geukes, S., Zwitserlood, P., & Christiansen, M. H. (2012). Processing multiple non-adjacent dependencies : Evidence from sequence learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2065-2076. https://doi.org/10.1098/rstb.2011.0414

Fiser, J., & Aslin, R. N. (2005). Encoding Multielement Scenes : Statistical Learning of Visual Feature Hierarchies. *Journal of Experimental Psychology: General*, *134*(4), 521-537. https://doi.org/10.1037/0096-3445.134.4.521

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. https://doi.org/10.1016/j.plrev.2014.04.005

Fitch, W. T., & Hauser, M. D. (2004). Computational Constraints on Syntactic Processing in a Nonhuman Primate. *Science*, *303*(5656), 377-380. https://doi.org/10.1126/science.1089401

Fitch, W. T., & Martins, M. J. D. (2014). Hierarchical processing in music, language, and action : Lashley revisited. *Annals of the New York Academy of Sciences*, *1316*(1), 87-104. https://doi.org/10.1111/nyas.12406

Forkstam, C., & Petersson, K. M. (2005). Towards an explicit account of implicit learning. *Current Opinion in Neurology*, *18*(4), 435-441. https://doi.org/10.1097/01.wco.0000171951.82995.c4

Franck, J., Rotondi, I., & Frauenfelder, U. H. (2016). Learning structure-dependent agreement in a hierarchical artificial grammar. *Journal of Memory and Language*, *87*, 84-104. https://doi.org/10.1016/j.jml.2015.11.003

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX : A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614-636. https://doi.org/10.1037/a0025255

Frensch, P. A., & Miner, C. S. (1994). Effects of presentation rate and individual differences in short-term memory capacity on an indirect measure of serial learning. *Memory & Cognition*, *22*(1), 95-110. https://doi.org/10.3758/BF03202765

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*. https://doi.org/10.3389/fnins.2016.00281

Geambaşu, A., Toron, L., Ravignani, A., & Levelt, C. C. (2020). Rhythmic Recursion? Human Sensitivity to a Lindenmayer Grammar with Self-similar Structure in a Musical Task. *Music & Science*, *3*, 205920432094661. https://doi.org/10.1177/2059204320946615

Gervain, J., de la Cruz-Pavía, I., & Gerken, L. (2020). Behavioral and imaging studies of infant artificial grammar learning. *Topics in Cognitive Science*, *12*, 815-827. https://doi.org/10.1111/tops.12400

Giroux, I., & Rey, A. (2009). Lexical and Sublexical Units in Speech Perception. *Cognitive Science*, *33*(2), 260-272. https://doi.org/10.1111/j.1551-6709.2009.01012.x

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation : Exploring the effects of context. *Cognition*, *112*(1), 21-54. https://doi.org/10.1016/j.cognition.2009.03.008

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178-186. https://doi.org/10.1016/S1364-6613(00)01467-4

Goschke, T. (1998). Implicit learning of perceptual and motor sequences : Evidence for independent learning systems. In *Handbook of implicit learning* (p. 401-444). Sage Publications, Inc.

Grama, I. C., Kerkhoff, A., & Wijnen, F. (2016). Gleaning Structure from Sound : The Role of Prosodic Contrast in Learning Non-adjacent Dependencies. *Journal of Psycholinguistic Research*, *45*(6), 1427-1449. https://doi.org/10.1007/s10936-016-9412-8

Hommel, B. (1994). Spontaneous decay of response-code activation. *Psychological Research*, *56*(4), 261-268. https://doi.org/10.1007/BF00419656

Honing, H., & Zuidema, W. (2014). Decomposing dendrophilia. *Physics of Life Reviews*, *11*(3), 375-376. https://doi.org/10.1016/j.plrev.2014.06.020

Huang, J., Dai, H., Ye, J., Zhu, C., Li, Y., & Liu, D. (2017). Impact of Response Stimulus Interval on Transfer of Non-local Dependent Rules in Implicit Learning : An ERP Investigation. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/article/10.3389/fpsyg.2017.02107

Huettel, S. A., Mack, P. B., & McCarthy, G. (2002). Perceiving patterns in random series : Dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience*, *5*(5), Article 5. https://doi.org/10.1038/nn841

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task : Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*(4), 658-680. https://doi.org/10.1037/0096-3445.130.4.658

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*, 188-196. https://doi.org/10.1037/h0056940

Isbilen, E. S., & Christiansen, M. H. (2022). Statistical Learning of Language : A Meta-Analysis Into 25 Years of Research. *Cognitive Science*, *46*(9), e13198. https://doi.org/10.1111/cogs.13198

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically Induced Chunking Recall : A Memory-Based Approach to Statistical Learning. *Cognitive Science*, *44*(7), e12848. https://doi.org/10.1111/cogs.12848

Jiménez, L., Méndez, A., Pasquali, A., Abrahamse, E., & Verwey, W. (2011). Chunking by colors : Assessing discrete learning in a continuous serial reaction-time task. *Acta Psychologica*, *137*(3), 318-329. https://doi.org/10.1016/j.actpsy.2011.03.013

Koch, I., & Hoffmann, J. (2000). Patterns, chunks, and hierarchies in serial reaction-time tasks. *Psychological Research*, *63*(1), 22-35. https://doi.org/10.1007/PL00008165

Koelsch, S. (2005). Neural substrates of processing syntax and semantics in music. *Current Opinion in Neurobiology*, *15*(2), 207-212. https://doi.org/10.1016/j.conb.2005.03.005

Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The Evolution of Rhythm Processing. *Trends in Cognitive Sciences*, *22*(10), 896-910. https://doi.org/10.1016/j.tics.2018.08.002

Kovács, Á. M., & Endress, A. D. (2014). Hierarchical Processing in Seven-Month-Old Infants. *Infancy*, *19*(4), 409-425. https://doi.org/10.1111/infa.12052

Krivochen, D., Phillips, B., & Saddy, J. (2018). *Classifying points in Lindenmayer systems : Transition probabilities and structure reconstruction (v. 1.1)*. https://doi.org/10.13140/RG.2.2.25719.88484

Kuhn, G., & Dienes, Z. (2006). Differences in the types of musical regularity learnt in incidental- and intentional-learning conditions. *Quarterly Journal of Experimental Psychology*, *59*(10), 1725-1744. https://doi.org/10.1080/17470210500438361

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package ‘lmertest’. *R package version*, *2*(0), 734.

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition*, *118*(2), 265-273. https://doi.org/10.1016/j.cognition.2010.11.011

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. https://doi.org/10.1080/20445911.2013.779247

Lashley, K. S. (1951). *The problem of serial order in behavior* (Vol. 21). Bobbs-Merrill Oxford, United Kingdom.

Levelt, W. J. M. (2020). On Empirical Methodology, Constraints, and Hierarchy in Artificial Grammar Learning. *Topics in Cognitive Science*, *12*(3), 942-956. https://doi.org/10.1111/tops.12441

Lewandowsky, S., & Oberauer, K. (2009). No evidence for temporal decay in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1545. https://doi.org/10.1037/a0017010

Lewis, S., & Phillips, C. (2015). Aligning Grammatical Theories and Language Processing Models. *Journal of Psycholinguistic Research*, *44*(1), 27-46. https://doi.org/10.1007/s10936-014-9329-z

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, *51*(1), 40-60. https://doi.org/10.3758/s13428-018-1076-x

Lindenmayer, A. (1968). Mathematical models for cellular interactions in development II. Simple and branching filaments with two-sided inputs. *Journal of Theoretical Biology*, *18*(3), 300-315. https://doi.org/10.1016/0022-5193(68)90080-5

Lynn, C. W., Kahn, A. E., Nyema, N., & Bassett, D. S. (2020). Abstract representations of events arise from mental errors in learning and memory. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15146-7

Maheu, M., Dehaene, S., & Meyniel, F. (2019). Brain signatures of a multiscale process of sequence learning in humans. *eLife*, *8*, e41541. https://doi.org/10.7554/eLife.41541

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. *Science*, *283*(5398), 77-80. https://doi.org/10.1126/science.283.5398.77

Martins, M. J. D. (2012). Distinctive signatures of recursion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2055–2064. https://doi.org/10.1098/rstb.2012.0097

Martins, M. J. D., Bianco, R., Sammler, D., & Villringer, A. (2019). Recursion in action : An fMRI study on the generation of new hierarchical levels in motor sequences. *Human Brain Mapping*, *40*(9), 2623-2638. https://doi.org/10.1002/hbm.24549

Martins, M. J. D., Fischmeister, F. P., Puig-Waldmüller, E., Oh, J., Geißler, A., Robinson, S., Fitch, W. T., & Beisteiner, R. (2014). Fractal image perception provides novel insights into hierarchical cognition. *NeuroImage*, *96*, 300-308. https://doi.org/10.1016/j.neuroimage.2014.03.064

Martins, M. J. D., Fischmeister, F. Ph. S., Gingras, B., Bianco, R., Puig-Waldmueller, E., Villringer, A., Fitch, W. T., & Beisteiner, R. (2020). Recursive music elucidates neural mechanisms supporting the generation and detection of melodic hierarchies. *Brain Structure and Function*, *225*(7), 1997-2015. https://doi.org/10.1007/s00429-020-02105-7

Martins, M. J. D., Gingras, B., Puig-Waldmueller, E., & Fitch, W. T. (2017). Cognitive representation of “musical fractals” : Processing hierarchy and recursion in the auditory domain. *Cognition*, *161*, 31-45. https://doi.org/10.1016/j.cognition.2017.01.001

Martins, M. J. D., Krause, C., Neville, D. A., Pino, D., Villringer, A., & Obrig, H. (2019). Recursive hierarchical embedding in vision is impaired by posterior middle temporal gyrus lesions. *Brain*, *142*(10), 3217-3229. https://doi.org/10.1093/brain/awz242

Martins, M. J. D., Muršič, Z., Oh, J., & Fitch, W. T. (2015). Representing visual recursion does not require verbal or motor resources. *Cognitive Psychology*, *77*, 20-41. https://doi.org/10.1016/j.cogpsych.2015.01.004

Maruyama, M., Pallier, C., Jobert, A., Sigman, M., & Dehaene, S. (2012). The cortical representation of simple mathematical expressions. *NeuroImage*, *61*(4), 1444-1460. https://doi.org/10.1016/j.neuroimage.2012.04.020

MATLAB. (2016). *Version 7.10.0 (R2016b)*. The MathWorks Inc.

McCarthy, G., & Donchin, E. (1981). A Metric for Thought : A Comparison of P300 Latency and Reaction Time. *Science*, *211*(4477), 77-80. https://doi.org/10.1126/science.7444452

McCauley, S. M., & Christiansen, M. H. (2014). Acquiring formulaic language : A computational model. *The Mental Lexicon*, *9*(3), 419-436. https://doi.org/10.1075/ml.9.3.03mcc

Meyniel, F., Maheu, M., & Dehaene, S. (2016). Human Inferences about Sequences : A Minimal Transition Probability Model. *PLOS Computational Biology*, *12*(12), e1005260. https://doi.org/10.1371/journal.pcbi.1005260

Monti, M. M., Parsons, L. M., & Osherson, D. N. (2012). Thought Beyond Language : Neural Dissociation of Algebra and Natural Language. *Psychological Science*, *23*(8), 914-922. https://doi.org/10.1177/0956797612437427

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2010). Learnability of Embedded Syntactic Structures Depends on Prosodic Cues. *Cognitive Science*, *34*(2), 338-349. https://doi.org/10.1111/j.1551-6709.2009.01093.x

Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical Implications of Articulatory Duration, Phonological Similarity, and Phonological Complexity in Verbal Working Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1353-1380. https://doi.org/10.1037/0278-7393.29.6.1353

Nakai, T., & Sakai, K. L. (2014). Neural Mechanisms Underlying the Computation of Hierarchical Tree Structures in Mathematics. *PLOS ONE*, *9*(11), e111439. https://doi.org/10.1371/journal.pone.0111439

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning : Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1-32. https://doi.org/10.1016/0010-0285(87)90002-8

Norman, E., Price, M. C., & Duff, S. C. (2006). Fringe consciousness in sequence learning : The influence of individual differences. *Consciousness and Cognition*, *15*(4), 723-760. https://doi.org/10.1016/j.concog.2005.06.003

Norman, E., Price, M. C., Duff, S. C., & Mentzoni, R. A. (2007). Gradations of awareness in a modified sequence learning task. *Consciousness and Cognition*, *16*(4), 809-837. https://doi.org/10.1016/j.concog.2007.02.004

Oberauer, K. (2013). The focus of attention in working memory—From metaphors to mechanisms. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00673

Oberauer, K. (2019). Working memory and attention—A conceptual analysis and review. *Journal of Cognition*, *2*. https://doi.org/10.5334/joc.58

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General*, *142*(2), 380-411. https://doi.org/10.1037/a0029588

Oberauer, K., & Lewandowsky, S. (2014). Further evidence against decay in working memory. *Journal of Memory and Language*, *73*, 15-30. https://doi.org/10.1016/j.jml.2014.02.003

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*(2), 225-237. https://doi.org/10.1016/j.jml.2005.02.011

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745-2750.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2 : Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-Driven Computations in Speech Processing. *Science*, *298*(5593), 604-607. https://doi.org/10.1126/science.1072901

Perruchet, P. (2019). What Mechanisms Underlie Implicit Statistical Learning? Transitional Probabilities Versus Chunks in Language Learning. *Topics in Cognitive Science*, *11*(3), 520-535. https://doi.org/10.1111/tops.12403

Perruchet, P., & Amorim, M.-A. (1992). Conscious knowledge and changes in performance in sequence learning : Evidence against dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 785-800. https://doi.org/10.1037/0278-7393.18.4.785

Perruchet, P., Bigand, E., & Benoit-Gonin, F. (1997). The emergence of explicit knowledge during the early phase of learning in sequential reaction time tasks. *Psychological Research*, *60*(1), 4-13. https://doi.org/10.1007/BF00419676

Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic Bulletin & Review*, *12*(2), 307-313. https://doi.org/10.3758/BF03196377

Perruchet, P., Robinet, V., & Lemaire, B. (2014). *U-Learn : Finding optimal coding units from unsegmented sequential databases*. http://leadserv.u-bourgogne.fr/~perruchet

Perruchet, P., & Vinter, A. (1998). PARSER : A Model for Word Segmentation. *Journal of Memory and Language*, *39*(2), 246-263. https://doi.org/10.1006/jmla.1998.2576

Planton, S., Kerkoerle, T. van, Abbih, L., Maheu, M., Meyniel, F., Sigman, M., Wang, L., Figueira, S., Romano, S., & Dehaene, S. (2021). A theory of memory for binary sequences : Evidence for a mental compression algorithm in humans. *PLOS Computational Biology*, *17*(1), e1008598. https://doi.org/10.1371/journal.pcbi.1008598

Pothos, E. (2010). An entropy model for artificial grammar learning. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00016

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, *133*(2), 227-244. https://doi.org/10.1037/0033-2909.133.2.227

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://support.pstnet.com/. (s. d.).

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Radulescu, S., Kotsolakou, A., Wijnen, F., Avrutin, S., & Grama, I. (2021). Fast but Not Furious. When Sped Up Bit Rate of Information Drives Rule Induction. *Frontiers in Psychology*, *12*, 4987. https://doi.org/10.3389/fpsyg.2021.661785

Radulescu, S., Wijnen, F., & Avrutin, S. (2019). Patterns Bit by Bit. An Entropy Model for Rule Induction. *Language Learning and Development*, 1-32. https://doi.org/10.1080/15475441.2019.1695620

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855-863. https://doi.org/10.1016/S0022-5371(67)80149-X

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests : Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 585-594. https://doi.org/10.1037/0278-7393.20.3.585

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories : The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30-54. https://doi.org/10.1016/j.cogpsych.2012.09.001

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020). Chapter 13 - More time for science : Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. In B. L. Parkin (Éd.), *Progress in Brain Research* (Vol. 253, p. 243-262). Elsevier. https://doi.org/10.1016/bs.pbr.2020.06.005

Ricker, T. J., Vergauwe, E., & Cowan, N. (2016). Decay Theory of Immediate Memory : From Brown (1958) to Today (2014). *Quarterly Journal of Experimental Psychology*, *69*(10), 1969-1995. https://doi.org/10.1080/17470218.2014.914546

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker : A MDL-Based Cognitive Model of Inductive Learning. *Cognitive Science*, *35*(7), 1352-1389. https://doi.org/10.1111/j.1551-6709.2011.01188.x

Saddy, J. D. (2009). Perceiving and processing recursion in formal grammars. *Recursion: Structural Complexity in Language and Cognition Conference at the University of Massachusetts (Amherst)*.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926-1928. https://doi.org/10.1126/science.274.5294.1926

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*(2), 229-242. https://doi.org/10.1007/s00221-003-1548-8

Savalia, T., Shukla, A., & Bapi, R. S. (2016). A Unified Theoretical Framework for Cognitive Sequencing. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/article/10.3389/fpsyg.2016.01821

Schiff, R., & Katan, P. (2014). Does complexity matter? Meta-analysis of learner performance in artificial grammar tasks. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/article/10.3389/fpsyg.2014.01084

Schmid, S., Saddy, D., & Franck, J. (2023a). Finding Hierarchical Structure in Binary Sequences : Evidence from Lindenmayer Grammar Learning. *Cognitive Science*, *47*(1), e13242. https://doi.org/10.1111/cogs.13242

Schmid, S., Saddy, D., & Franck, J. (2023b). *Uncovering hierarchical structure through statistical dependency*. PsyArXiv [Preprint]. https://doi.org/10.31234/osf.io/mre4h

Schwarb, H., & Schumacher, E. H. (2012). Generalized lessons about sequence learning from the study of the serial reaction time task. *Advances in Cognitive Psychology*, *8*(2), 165-178. https://doi.org/10.5709/acp-0113-1

Shanks, D. R., & Johnstone, T. (1999). Evaluating the relationship between explicit and implicit knowledge in a sequential reaction time task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1435-1451. https://doi.org/10.1037/0278-7393.25.6.1435

Shanks, D. R., Wilkinson, L., & Channon, S. (2003). Relationship between priming and recognition in deterministic and probabilistic sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 248-261. https://doi.org/10.1037/0278-7393.29.2.248

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shirley, E. J. (2014). *Representing and remembering Lindenmayer-grammars* [Ph.D., University of Reading]. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.658878

Simon, H. A. (1962). An information processing theory of intellectual development. *Monographs of the Society for Research in Child Development*, 150-161.

Slone, L., & Johnson, S. P. (2015). Statistical and Chunking Processes in Adults' Visual Sequence Learning. *CogSci*.

Slone, L. K., & Johnson, S. P. (2018). When learning goes beyond statistics : Infants represent visual sequences in terms of chunks. *Cognition*, *178*, 92-102. https://doi.org/10.1016/j.cognition.2018.05.016

Soetens, E., Boer, L. C., & Hueting, J. E. (1985). Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 598. https://doi.org/10.1037/0096-1523.11.5.598

Soetens, E., Melis, A., & Notebaert, W. (2004). Sequence learning and sequential effects. *Psychological Research*, *69*(1), 124-137. https://doi.org/10.1007/s00426-003-0163-4

Stadler, M. A. (1995). Role of attention in implicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 674-685. https://doi.org/10.1037/0278-7393.21.3.674

Sternberg, S. (1969). Memory-Scanning : Mental Processes Revealed by Reaction-Time Experiments. *American Scientist*, *57*(4), 421-457. https://www.jstor.org/stable/27828738

ten Cate, C., Gervain, J., Levelt, C. C., Petkov, C. I., & Zuidema, W. (2020). Editors' Review and Introduction : Learning Grammatical Structures: Developmental, Cross-Species, and Computational Approaches. *Topics in Cognitive Science*, *12*(3), 804-814. https://doi.org/10.1111/tops.12493

Thiessen, E. D., & Erickson, L. C. (2013). Beyond Word Segmentation : A Two- Process Account of Statistical Learning. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721413476035

Trotter, A. S., Monaghan, P., Beckers, G. J. L., & Christiansen, M. H. (2020). Exploring Variation Between Artificial Grammar Learning Experiments : Outlining a Meta-Analysis Approach. *Topics in Cognitive Science*, *12*(3), 875-893. https://doi.org/10.1111/tops.12454

Uddén, J., Martins, M. de J. D., Zuidema, W., & Fitch, W. T. (2020). Hierarchical Structure in Sequence Processing : How to Measure It and Determine Its Neural Implementation. *Topics in Cognitive Science*, *12*(3), 910-924. https://doi.org/10.1111/tops.12442

van den Bos, E., Christiansen, M. H., & Misyak, J. B. (2012). Statistical learning of probabilistic nonadjacent dependencies by multiple-cue integration. *Journal of Memory and Language*, *67*(4), 507-520. https://doi.org/10.1016/j.jml.2012.07.008

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance : A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653-673. https://doi.org/10.3758/s13428-016-0721-5

Vandierendonck, A. (2018). Further Tests of the Utility of Integrated Speed-Accuracy Measures in Task Switching. *Journal of Cognition*, *1*(1), 8. https://doi.org/10.5334/joc.6

Vender, M., Krivochen, D. G., Compostella, A., Phillips, B., Delfitto, D., & Saddy, D. (2020). Disentangling sequential from hierarchical learning in Artificial Grammar Learning : Evidence from a modified Simon Task. *PLOS ONE*, *15*(5), e0232687. https://doi.org/10.1371/journal.pone.0232687

Vender, M., Krivochen, D. G., Phillips, B., Saddy, D., & Delfitto, D. (2019). Implicit Learning, Bilingualism, and Dyslexia : Insights From a Study Assessing AGL With a Modified Simon Task. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01647

Verwey, W. B., & Dronkers, W. J. (2019). Skill in discrete keying sequences is execution rate specific. *Psychological research*. https://doi.org/10.1007/s00426-017-0967-2

Verwey, W. B., & Wright, D. L. (2014). Learning a keying sequence you never executed : Evidence for independent associative and motor chunk learning. *Acta Psychologica*, *151*, 24-31. https://doi.org/10.1016/j.actpsy.2014.05.017

Vitányi, P. M. B., & Walker, A. (1978). Stable string languages of lindenmayer systems. *Information and Control*, *37*(2), 134-149. https://doi.org/10.1016/S0019-9958(78)90483-7

Wilkinson, L., & Shanks, D. R. (2004). Intentional Control and Implicit Sequence Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 354-369. https://doi.org/10.1037/0278-7393.30.2.354

Willingham, D. B., Greenberg, A. R., & Thomas, R. C. (1997). Response-to-stimulus interval does not affect implicit motor sequence learning, but does affect performance. *Memory & Cognition*, *25*(4), 534-542. https://doi.org/10.3758/BF03201128

7. Supplementary materials

7.1 Supplementary materials for Chapter 2

S1 – Transition patterns for the constituent [01101]

To evaluate if the slowdown at the second transition of the constituent [01101] found at the group level was due to a subset of participants, we considered each of the 4 transition for each individual. This analysis was done only on the constituent [01101] when it occurred in non-ambiguous structural context. In order to increase statistical power, we considered block 4 and 5 jointly.

We ran 4 linear models (one for each transition) on reaction times. The factor *Position* had two modality (before, after), “before” coded for the points that was before the transition and “after” coded for the point after the transition. Each models had reaction times as dependent variable and as predictor the factors *Participants* and the interaction *Participants* Position*. The factor *Position* was entered only in the interaction term in order to compare the effect of position for the same individual and not across individuals. The modality “before” of the factor *Position* was always set as the intercept, thus, a “-” sign before the interaction coefficients indicate a diminution of RTs. We considered in the analysis only trials in which a correct response was given. Since the variable *Participants* was used as factor, the p-value of the main effects and the global interaction are overestimated, we thus did not interpret them and we report only the interaction coefficient of each participants.

Mean difference of RTs (ms) and Standard error for Transition 1 of the constituent [01101] in non-ambiguous structural context by Position and Participants

Participants	*Position before - after*	*Std.Error*	*T-value*
1	-101	27	-3.75 ***
2	-157	24.4	-6.42 ***
3	-111	27.4	-4.07 ***
4	-99	25.1	-3.95 ***

5	-94	28.9	-3.25 **
6	-147	24.8	-5.93 ***
7	-75	30.5	-2.47 *
8	-131	34.1	-3.84 ***
9	-56	27	-2.06 *
10	-38	29.4	-1.29
11	-274	26.6	-10.31 ***
12	-130	30.5	-4.25 ***
13	-184	25.8	-7.12 ***
14	-111	27	-4.1 ***
15	-247	28.9	-8.58 ***
16	-141	25.1	-5.61 ***
17	-147	31.8	-4.63 ***
18	-80	24.4	-3.27 **
19	-123	31.8	-3.87 ***
20	-117	29.9	-3.89 ***
21	-123	25.1	-4.89 ***
22	-163	29.4	-5.53 ***
23	-113	27	-4.18 ***
24	-149	25.1	-5.94 ***
25	-68	24.8	-2.76 **
26	-114	25.1	-4.53 ***
27	-159	26.2	-6.06 ***
28	-166	30.5	-5.44 ***
29	2	35	0.06
30	-103	25.4	-4.03 ***
31	-136	24.4	-5.57 ***
32	-81	34.1	-2.37 *
33	-25	25.1	-0.98
34	-129	25.4	-5.08 ***
35	-98	24.4	-4.01 ***
36	-217	28.4	-7.67 ***
37	-93	27.9	-3.35 ***
38	-124	25.4	-4.89 ***
39	-130	27	-4.82 ***
40	-57	28.9	-1.97 *
41	-106	25.1	-4.23 ***
42	-153	25.1	-6.09 ***
43	-67	27.9	-2.42 *
44	-67	31.2	-2.15 *
45	-137	27.9	-4.91 ***
46	-159	24.4	-6.52 ***
47	-45	25.8	-1.73 .
48	-135	24.1	-5.6 ***
49	-199	27.9	-7.14 ***
50	-120	28.9	-4.17 ***
51	-122	28.9	-4.22 ***
52	-42	26.2	-1.6
53	-130	27	-4.81 ***
54	-85	23.8	-3.58 ***

55	-200	31.8	-6.29 ***
56	-248	30.5	-8.11 ***
57	-53	23.8	-2.22 *
58	-255	27.4	-9.3 ***
59	-156	25.8	-6.04 ***
60	-65	27.4	-2.35 *
61	-92	24.8	-3.71 ***
62	-116	27	-4.29 ***
63	-81	27.9	-2.91 **
64	-128	27	-4.76 ***
65	-123	32.6	-3.78 ***
66	-128	26.2	-4.88 ***
67	-124	29.4	-4.21 ***
68	-113	27.4	-4.11 ***
69	-82	25.8	-3.18 **
70	-90	31.8	-2.84 **
71	-98	35	-2.79 **
72	-147	31.8	-4.62 ***
73	-150	27	-5.57 ***
74	-117	27.4	-4.28 ***
75	-5	28.9	-0.16
76	-65	25.1	-2.58 **
77	-98	27.4	-3.59 ***
78	-171	25.1	-6.83 ***
79	-194	25.4	-7.63 ***
80	-94	29.4	-3.21 **
81	-82	25.1	-3.25 **
82	-113	29.4	-3.83 ***
83	-99	25.8	-3.84 ***
84	-176	27	-6.51 ***
85	-147	29.9	-4.9 ***
86	-158	26.6	-5.96 ***
87	-94	30.5	-3.07 **
88	-178	25.8	-6.92 ***
89	-118	27.4	-4.32 ***
90	-224	31.8	-7.05 ***
91	-293	28.9	-10.14 ***
92	-117	27.9	-4.19 ***
93	-92	26.6	-3.44 ***
94	-121	25.1	-4.82 ***
95	-180	24.8	-7.26 ***
96	-55	23.8	-2.3 *
97	-17	25.4	-0.67
98	-85	27	-3.15 **
99	-115	25.1	-4.57 ***
100	-126	26.6	-4.73 ***
101	-102	27.4	-3.74 ***
102	-169	27	-6.28 ***
103	-158	30.5	-5.18 ***
104	-41	25.4	-1.6

105	-84	29.4	-2.85 **
106	-155	24.4	-6.33 ***
107	-80	28.4	-2.82 **
108	-112	28.9	-3.87 ***
109	-194	28.4	-6.85 ***
110	-82	27	-3.03 **
111	-133	27.4	-4.84 ***
112	-125	33.3	-3.75 ***
113	-82	24.4	-3.35 ***
114	-52	26.6	-1.94 .
115	-213	25.8	-8.26 ***
116	-101	28.9	-3.52 ***
117	-131	28.9	-4.54 ***
118	-106	29.4	-3.6 ***
119	-96	31.2	-3.07 **
120	-183	26.6	-6.9 ***
121	-170	25.8	-6.59 ***
122	-109	27	-4.04 ***
123	-45	24.8	-1.8 .
124	-222	27.9	-7.96 ***
125	-104	25.8	-4.04 ***
126	-120	25.4	-4.71 ***
127	-149	25.1	-5.93 ***
128	-164	25.4	-6.46 ***
129	-198	25.8	-7.66 ***
130	-99	27.4	-3.62 ***
131	-210	25.1	-8.35 ***
132	-108	30.5	-3.54 ***
133	-141	25.8	-5.47 ***
134	-135	28.9	-4.68 ***
135	-108	28.9	-3.75 ***
136	-223	26.2	-8.5 ***
137	-91	24.4	-3.73 ***
138	-103	27	-3.8 ***
139	-276	34.1	-8.07 ***
140	-123	24.4	-5.01 ***
141	-90	24.8	-3.63 ***
142	-106	24.4	-4.35 ***
143	-201	27	-7.44 ***
144	-45	27	-1.68 .
145	-164	32.6	-5.03 ***
146	-139	25.1	-5.56 ***
147	-88	25.1	-3.5 ***
148	-26	25.4	-1.01
149	-85	26.2	-3.23 **
150	-90	26.6	-3.39 ***
151	-148	27.4	-5.39 ***
152	-107	24.1	-4.43 ***
153	-221	25.1	-8.81 ***
154	-100	27.4	-3.66 ***

155	-96	25.1	-3.84 ***
156	-176	25.4	-6.93 ***
157	-172	24.8	-6.95 ***
158	-99	24.4	-4.07 ***
159	-150	26.2	-5.74 ***

Note. .p < .1, *p < .05, **p < .01, ***p < .001

Mean difference of RTs (ms) and Standard error for Transition 2 of the constituent [01101] in non-ambiguous structural context by Position and Participants

Participants	*Position before - after*	*Std.Error*	*T-value*
1	88	25.9	3.39***
2	51	23.5	2.17*
3	17	26.3	0.63
4	78	24.1	3.25**
5	64	27.7	2.31*
6	144	23.8	6.07***
7	64	29.3	2.18*
8	95	32.8	2.89**
9	5	25.9	0.18
10	78	28.2	2.77**
11	234	25.5	9.18***
12	147	29.3	5.01***
13	42	24.8	1.68.
14	36	25.9	1.38
15	177	27.7	6.40***
16	46	24.1	1.89.
17	120	30.5	3.94***
18	103	23.5	4.40***
19	17	30.5	0.54
20	109	28.7	3.78***
21	52	24.1	2.14*
22	126	28.2	4.47***
23	94	25.9	3.62***
24	106	24.1	4.40***
25	-26	23.8	-1.09
26	58	24.1	2.40*
27	118	25.1	4.7***
28	124	29.3	4.22***
29	-42	33.6	-1.24
30	45	24.4	1.84.
31	36	23.5	1.52
32	66	32.8	2.02*
33	48	24.1	1.99*
34	17	24.4	0.69
35	36	23.5	1.52
36	193	27.2	7.08***
37	86	26.7	3.20**
38	95	24.4	3.89***
39	79	25.9	3.04**
40	71	27.7	2.56*
41	46	24.1	1.91.
42	121	24.1	5.03***
43	30	26.7	1.11
44	55	29.9	1.83.
45	-7	26.7	-0.25
46	118	23.5	5.03***
47	22	24.8	0.86
48	53	23.2	2.29*
49	80	26.7	2.98**

50	51	27.7	1.82.
51	94	27.7	3.39***
52	10	25.1	0.38
53	50	25.9	1.94.
54	23	22.9	1.00
55	123	30.5	4.04***
56	131	29.3	4.47***
57	29	22.9	1.25
58	216	26.3	8.21***
59	159	24.8	6.43***
60	17	26.3	0.63
61	77	23.8	3.24**
62	67	25.9	2.57*
63	28	26.7	1.06
64	110	25.9	4.24***
65	138	31.2	4.41***
66	105	25.1	4.19***
67	129	28.2	4.56***
68	116	26.3	4.39***
69	37	24.8	1.48
70	91	30.5	2.98**
71	68	33.6	2.01*
72	147	30.5	4.81***
73	39	25.9	1.50
74	74	26.3	2.79**
75	-19	27.7	-0.69
76	8	24.1	0.32
77	98	26.3	3.73***
78	6	24.1	0.24
79	127	24.4	5.20***
80	47	28.2	1.67.
81	32	24.1	1.32
82	79	28.2	2.80**
83	56	24.8	2.26*
84	151	25.9	5.81***
85	190	28.7	6.60***
86	130	25.5	5.09***
87	84	29.3	2.88**
88	152	24.8	6.13***
89	61	26.3	2.32*
90	218	30.5	7.14***
91	252	27.7	9.10***
92	118	26.7	4.42***
93	98	25.5	3.84***
94	78	24.1	3.23**
95	128	23.8	5.36***
96	31	22.9	1.34
97	21	24.4	0.84
98	130	25.9	5.03***
99	94	24.1	3.88***
100	69	25.5	2.70**
101	116	26.3	4.39***
102	143	25.9	5.51***
103	102	29.3	3.49***
104	37	24.4	1.52
105	89	28.2	3.14**
106	114	23.5	4.87***
107	114	27.2	4.20***
108	89	27.7	3.19**
109	106	27.2	3.90***
110	93	25.9	3.60***
111	69	26.3	2.63**
112	51	32	1.58

113	34	23.5	1.44
114	54	25.5	2.12*
115	134	24.8	5.41***
116	110	27.7	3.96***
117	66	27.7	2.39*
118	109	28.2	3.85***
119	94	29.9	3.15**
120	126	25.5	4.93***
121	90	24.8	3.61***
122	52	25.9	2*
123	10	23.8	0.44
124	79	26.7	2.96**
125	49	24.8	1.97*
126	121	24.4	4.97***
127	52	24.1	2.16*
128	57	24.4	2.31*
129	109	24.8	4.40***
130	46	26.3	1.75.
131	76	24.1	3.17**
132	99	29.3	3.37***
133	63	24.8	2.52*
134	91	27.7	3.27**
135	38	27.7	1.36
136	125	25.1	4.95***
137	39	23.5	1.68.
138	86	25.9	3.33***
139	198	32.8	6.03***
140	82	23.5	3.50***
141	44	23.8	1.84.
142	76	23.5	3.22**
143	145	25.9	5.60***
144	-8	25.9	-0.31
145	97	31.2	3.10**
146	98	24.1	4.08***
147	63	24.1	2.59**
148	-14	24.4	-0.57
149	56	25.1	2.21*
150	-8	25.5	-0.33
151	96	26.3	3.65***
152	10	23.2	0.44
153	154	24.1	6.38***
154	31	26.3	1.18
155	61	24.1	2.54*
156	170	24.4	6.98***
157	118	23.8	4.95***
158	70	23.5	3.00**
159	190	25.1	7.57***

Note. .$p < .1$, *$p < .05$, **$p < .01$, ***$p < .001$

Mean difference of RTs (ms) and Standard error for Transition 3 of the constituent [01101] in non-ambiguous structural context by Position and Participants

Participants	*Position before - after*	*Std.Error*	*T-value*
1	-92	26	-3.53***
2	-44	23.5	-1.89.
3	54	26.4	2.06*
4	-99	24.2	-4.10***
5	-105	27.8	-3.78***
6	-66	23.9	-2.77**
7	-3	29.4	-0.10
8	14	32.9	0.43

9	-30	26	-1.16
10	-46	28.3	-1.63
11	-50	25.6	-1.94.
12	-49	29.4	-1.68.
13	-5	24.9	-0.19
14	-10	26	-0.36
15	-47	27.8	-1.70.
16	-72	24.2	-2.99**
17	-40	30.7	-1.30
18	-89	23.5	-3.78***
19	-10	30.7	-0.33
20	-37	28.8	-1.27
21	-22	24.2	-0.92
22	67	28.3	2.36*
23	-69	26	-2.63**
24	-54	24.2	-2.22*
25	24	23.9	1.00
26	-2	24.2	-0.08
27	-35	25.2	-1.37
28	-12	29.4	-0.4
29	32	33.7	0.95
30	-17	24.5	-0.68
31	-13	23.5	-0.54
32	-3	32.9	-0.10
33	-55	24.2	-2.28*
34	37	24.5	1.52
35	6	23.5	0.23
36	-121	27.3	-4.42***
37	-48	26.8	-1.80.
38	-65	24.5	-2.64**
39	-15	26	-0.59
40	-71	27.8	-2.55*
41	-22	24.2	-0.92
42	-94	24.2	-3.88***
43	-7	26.8	-0.25
44	-20	30	-0.66
45	34	26.8	1.25
46	-124	23.5	-5.26***
47	-43	24.9	-1.71.
48	-63	23.2	-2.71**
49	-36	26.8	-1.32
50	-13	27.8	-0.47
51	-79	27.8	-2.82**
52	-38	25.2	-1.48
53	-17	26	-0.65
54	-28	23	-1.21
55	-29	30.7	-0.93
56	-63	29.4	-2.14*
57	-26	23	-1.14
58	-58	26.4	-2.19*
59	-29	24.9	-1.17
60	6	26.4	0.22
61	-56	23.9	-2.34*
62	-20	26	-0.76
63	-115	26.8	-4.28***
64	-59	26	-2.27*
65	-35	31.3	-1.10
66	-66	25.2	-2.59**
67	-88	28.3	-3.09**
68	-60	26.4	-2.28*
69	-40	24.9	-1.59
70	-40	30.7	-1.31
71	44	33.7	1.31

72	-64	30.7	-2.07*
73	15	26	0.56
74	27	26.4	1.01
75	-56	27.8	-2.00*
76	12	24.2	0.48
77	-13	26.4	-0.47
78	37	24.2	1.52
79	-103	24.5	-4.20***
80	-35	28.3	-1.23
81	-37	24.2	-1.53
82	-10	28.3	-0.36
83	-14	24.9	-0.56
84	-43	26	-1.65.
85	-65	28.8	-2.24*
86	-117	25.6	-4.56***
87	-42	29.4	-1.44
88	-59	24.9	-2.36*
89	-60	26.4	-2.28*
90	-84	30.7	-2.72**
91	-24	27.8	-0.84
92	-53	26.8	-1.97*
93	-151	25.6	-5.90***
94	-52	24.2	-2.13*
95	-66	23.9	-2.74**
96	-62	23	-2.71**
97	-38	24.5	-1.53
98	-74	26	-2.86**
99	-50	24.2	-2.05*
100	7	25.6	0.27
101	-17	26.4	-0.66
102	-99	26	-3.82***
103	-28	29.4	-0.95
104	-43	24.5	-1.73.
105	-50	28.3	-1.77.
106	-50	23.5	-2.12*
107	-33	27.3	-1.20
108	-58	27.8	-2.07*
109	17	27.3	0.61
110	-54	26	-2.09*
111	-30	26.4	-1.15
112	-35	32.1	-1.09
113	-30	23.5	-1.25
114	-52	25.6	-2.01*
115	-18	24.9	-0.72
116	-40	27.8	-1.42
117	5	27.8	0.19
118	-55	28.3	-1.93.
119	-41	30	-1.36
120	-43	25.6	-1.66.
121	-77	24.9	-3.09**
122	-3	26	-0.11
123	-36	23.9	-1.52
124	41	26.8	1.53
125	9	24.9	0.35
126	-78	24.5	-3.18**
127	-61	24.2	-2.51*
128	46	24.5	1.88.
129	-92	24.9	-3.70***
130	-28	26.4	-1.06
131	-11	24.2	-0.46
132	-55	29.4	-1.86.
133	-13	24.9	-0.51
134	-36	27.8	-1.29

135	-16	27.8	-0.59
136	-134	25.2	-5.32***
137	-31	23.5	-1.31
138	-34	26	-1.29
139	-5	32.9	-0.15
140	-48	23.5	-2.04*
141	-15	23.9	-0.61
142	-66	23.5	-2.79**
143	-73	26	-2.80**
144	-92	26	-3.55***
145	-100	31.3	-3.18**
146	-74	24.2	-3.04**
147	-56	24.2	-2.32*
148	-56	24.5	-2.27*
149	-1	25.2	-0.03
150	-93	25.6	-3.62***
151	-102	26.4	-3.86***
152	-48	23.2	-2.04*
153	-88	24.2	-3.64***
154	-3	26.4	-0.09
155	-31	24.2	-1.28
156	-104	24.5	-4.25***
157	-59	23.9	-2.46*
158	-12	23.5	-0.50
159	-109	25.2	-4.32***

Note. .*p* < .1, **p* < .05, ***p* < .01, ****p* < .001

Mean difference of RTs (ms) and Standard error for Transition 4 of the constituent [01101] in non-ambiguous structural context by Position and Participants

Participants	*Position before - after*	*Std.Error*	*T-value*
1	11	25.6	0.42
2	42	23.2	1.78.
3	-109	26	-4.17***
4	34	23.8	1.43
5	42	27.4	1.53
6	-49	23.5	-2.07*
7	-58	29	-2.01*
8	-108	32.4	-3.32***
9	30	25.6	1.17
10	-9	27.9	-0.34
11	-252	25.2	-9.97***
12	-56	29	-1.94.
13	16	24.5	0.65
14	13	25.6	0.49
15	-132	27.4	-4.81***
16	36	23.8	1.52
17	-76	30.2	-2.50*
18	-18	23.2	-0.75
19	-24	30.2	-0.79
20	-58	28.4	-2.05*
21	-48	23.8	-2.01*
22	-180	27.9	-6.43***
23	-12	25.6	-0.48
24	-42	23.8	-1.77.
25	30	23.5	1.26
26	-32	23.8	-1.35
27	-74	24.9	-2.97**
28	-84	29	-2.9**
29	-22	33.3	-0.67
30	8	24.2	0.34

31	43	23.2	1.85.
32	-75	32.4	-2.30*
33	6	23.8	0.25
34	18	24.2	0.74
35	13	23.2	0.54
36	-30	26.9	-1.10
37	-62	26.5	-2.33*
38	30	24.2	1.24
39	-72	25.6	-2.80**
40	4	27.4	0.16
41	-10	23.8	-0.40
42	-43	23.8	-1.81.
43	-39	26.5	-1.49
44	-75	29.6	-2.52*
45	-25	26.5	-0.93
46	42	23.2	1.79.
47	6	24.5	0.26
48	25	22.9	1.08
49	-87	26.5	-3.28**
50	-19	27.4	-0.70
51	5	27.4	0.19
52	20	24.9	0.81
53	20	25.6	0.78
54	14	22.6	0.6
55	-127	30.2	-4.19***
56	-9	29	-0.29
57	-7	22.6	-0.31
58	-163	26	-6.26***
59	-99	24.5	-4.05***
60	-65	26	-2.49*
61	-48	23.5	-2.02*
62	-27	25.6	-1.05
63	3	26.5	0.10
64	-45	25.6	-1.75.
65	-89	30.9	-2.87**
66	16	24.9	0.63
67	-95	27.9	-3.42***
68	-68	26	-2.61**
69	-5	24.5	-0.18
70	-47	30.2	-1.54
71	-81	33.3	-2.42*
72	-91	30.2	-3.01**
73	-36	25.6	-1.40
74	-104	26	-3.98***
75	26	27.4	0.96
76	-6	23.8	-0.26
77	-61	26	-2.33*
78	4	23.8	0.16
79	-6	24.2	-0.25
80	21	27.9	0.74
81	9	23.8	0.38
82	-40	27.9	-1.42
83	-10	24.5	-0.39
84	-67	25.6	-2.63**
85	-145	28.4	-5.08***
86	20	25.2	0.80
87	-44	29	-1.52
88	-56	24.5	-2.28*
89	-4	26	-0.17
90	-102	30.2	-3.38***
91	-279	27.4	-10.16***
92	-68	26.5	-2.57*
93	114	25.2	4.52***

94	-18	23.8	-0.76
95	-39	23.5	-1.66.
96	34	22.6	1.50
97	1	24.2	0.05
98	-95	25.6	-3.70***
99	-57	23.8	-2.4*
100	-50	25.2	-1.97*
101	-80	26	-3.08**
102	-35	25.6	-1.34
103	-40	29	-1.37
104	31	24.2	1.29
105	-98	27.9	-3.49***
106	-34	23.2	-1.47
107	-65	26.9	-2.41*
108	-30	27.4	-1.08
109	-74	26.9	-2.73**
110	-14	25.6	-0.55
111	0	26	0.01
112	-12	31.6	-0.36
113	-4	23.2	-0.18
114	1	25.2	0.02
115	-189	24.5	-7.69***
116	-38	27.4	-1.39
117	-80	27.4	-2.93**
118	-66	27.9	-2.37*
119	-65	29.6	-2.19*
120	-57	25.2	-2.25*
121	-9	24.5	-0.37
122	-49	25.6	-1.90.
123	10	23.5	0.42
124	-125	26.5	-4.72***
125	-51	24.5	-2.08*
126	-31	24.2	-1.27
127	6	23.8	0.23
128	-107	24.2	-4.41***
129	-5	24.5	-0.20
130	5	26	0.20
131	-12	23.8	-0.48
132	-64	29	-2.2*
133	-47	24.5	-1.92.
134	-36	27.4	-1.32
135	21	27.4	0.75
136	85	24.9	3.43***
137	-11	23.2	-0.48
138	13	25.6	0.52
139	-151	32.4	-4.66***
140	-35	23.2	-1.51
141	-28	23.5	-1.19
142	-19	23.2	-0.80
143	-67	25.6	-2.60**
144	94	25.6	3.67***
145	67	30.9	2.15*
146	-10	23.8	-0.42
147	-5	23.8	-0.21
148	58	24.2	2.40*
149	-47	24.9	-1.89.
150	83	25.2	3.29***
151	36	26	1.38
152	45	22.9	1.95.
153	-64	23.8	-2.68**
154	-14	26	-0.53
155	-27	23.8	-1.13
156	-106	24.2	-4.36***

157	-54	23.5	-2.28*
158	-49	23.2	-2.09*
159	-47	24.9	-1.88.

Note. $.p < .1$, $*p < .05$, $**p < .01$, $***p < .001$

S2 - Data reported in Chapter 2

Deidentified associated with Chapter 2 can be found electronically at https://osf.io/8n9he/?view_only=6f4b42d8e0d7429a984d9a8ff96ad4ba

7.2 Supplementary materials for Chapter 3

S3 - Data reported in Chapter 3

Deidentified data collected in Experiments 1 and 2 of Chapter 3 are posted at https://osf.io/jauq3/?view_only=86efbe9dfa66487284da6d7039de3643.

7.3 Supplementary materials for Chapter 5

S4- List of n-grams used in the PARSER simulation

Each list contains all possible n-grams in Fib for a given length. Since PARSER can track up to 10 lists simultaneously, we have considered all possible n-grams up to a length of 10.

Lenght	**Identity**
1-grams	1
	0
2-grams	01
	10
	11
3-grams	110
	101
	011
	010
4-grams	1101
	1011
	1010
	0110
	0101

Lenght	**Identity**
5-grams	11011
	11010
	10110
	10101
	01101
	01011
6-grams	110110
	110101
	101101
	101011
	011011
	011010
	010110

Lenght	**Identity**
7-grams	1101101
	1101011
	1011011
	1011010
	1010110
	0110101
	0110110
	0101101
8-grams	11011010
	11010110
	10110110
	10110101
	10101101
	01101101
	01101011
	01011011
	01011010
9-grams	110110101
	110101101
	101101101
	101101011
	101011011
	101011010
	011011010
	011010110
	010110110
	010110101

Lenght	**Identity**
10-grams	1101101011
	1101011010
	1101011011
	1011011010
	1011010110
	1010110110
	1010110101
	0110110101
	0110101101
	0101101101
	0101101011