UTRECHT UNIVERSITY

MASTER THESIS

FACULTY OF SCIENCE – ARTIFICIAL INTELLIGENCE

Capacity Modeling in Animal AGL Studies

Author:
Bror Emil Baklund Krogsrud
S.n 5978602

Supervisors:
Dr. Gabriël J.L. Beckers
Dr. Frans W. Adriaans

November 22, 2018

Universiteit Utrecht

Abstract

Artificial grammar learning (AGL) is an experimental paradigm used to investigate the processes that underlie language learning. Participants are trained on a set of strings generated from a grammar before they are being tested on their ability to generalize their learning to novel strings generated from the same grammar. From early AGL experiments in humans it was concluded that the participants learn to represent and encode the grammar in the form of rules. Later, this claim has been challenged by a number of other possible learning strategies and mechanisms that can account for what is being learned, and how it's being learned. Work on humans has so far not yielded a consensus as to what the mechanisms driving AGL are. However, it has yield a number of computational models able to account for the performance reported in human studies. Since recently, the AGL paradigm has also been used to investigate potential grammatical competence in non-human animals. But so far, animal AGL studies have not embraced the possibility of other mechanisms than rule-learning, leaving computational models able to account for the phenomena in humans, unexplored. In this thesis, I first review the animal AGL literature and find that the tacit assumption in many studies is that rule-learning is the only possible mechanism of AGL, leading to experimental designs that do not control for other, potentially simpler, explanations. To challenge this bias, I implement and apply a computational model (PARSER) for word-segmentation in human infants, to experimental data from animal AGL experiments. In silico replication of animal AGL experiments show that PARSER is indeed capable of accounting for the AGL performance reported in animal studies. From these experimental results, it is concluded that mechanisms other than rule-learning are equally, or more likely to explain the performance of animals in artificial grammar learning tasks.

Table of contents

1 Introduction

This chapter has the purpose of putting forward the focus of the research conducted in this thesis. First I present the motivation for the research project, then two research questions are put forward before the outline of the remainder of the thesis is presented.

1.1. Motivation

A fundamental question in cognitive science is whether the capacity to use language is unique to humans. The linguistic capacities of humans give us the ability to understand and generate language through a number of sensory modalities, such as structured vocalization, written text, or braille, as well as to extract semantic values from these structures. Mapping of the cognitive capacities required for human-level language processing remains an active area of investigation and is commonly addressed through comparative work on evolutionarily conserved mechanisms in humans and non-human animals[1]. The processing of grammatical rules, constituting what the syntax of human languages is, is a core component of the human language faculty[2]. Cross-species comparative work has shown essential differences in how species processes such grammatical rules[3]. Recent experimental studies utilizing the artificial grammar learning (AGL) paradigm in non-human animals reports grammatical capacities in a wide variety of species[4]–[7], although rarely at the level of human natural grammar.

AGL is a research paradigm initially intended for the study of human rule learning[8] but has later been adopted for the study of implicit learning [9]. More recently, the paradigm is being used in a wide array of inquiries such as assessing the syntactic capabilities of animals [4] and the learning abilities of amnesic patients [10]. In experiments that make use of this paradigm participants are familiarized with a set of strings which, unbeknownst to them, are made from an artificial grammar. Following the familiarization phase, the participants are tested on their ability to discriminate between strings that are consistent with the rules of the grammar and strings that are inconsistent with the grammar. The experimenter has the knowledge of both the grammar and the full language the grammar can produce. Logically this yields that an experimenter can feed the participant a subset of the strings from a grammar produced language, to investigate their ability to infer, represent and use the grammar [4]. It becomes important to note that in all AGL studies that do not use neuroimaging techniques such as fMRI and EEG, grammatical inference performance and representation estimation are assumed from the participants' usage of the grammar. It is, for this reason, the paradigm can accommodate both human and non-human animal research.

In the field of human AGL research, it has become more and more realized that grammar may not be the only thing learned in AGL experiments. Perceived performance of grammar inference may not correctly represent the grammar inference itself, but rather the participants' ability to discriminate

strings based on learned stimulus fragments[11] or other similar mechanisms. Naturally, accounting for such a problem becomes an intrinsically difficult task if the goal is to keep the claim of grammaticality. The paradigm is undergoing a shift from addressing the question *if* participants can learn grammars, towards questions asking *what* they are learning and *how* they are learning it. Several theoretical accounts detailing mechanisms of *what* and *how* have been proposed, but there exists no general agreement about the importance and interplay between them[12]. However, as a natural consequence of having detailed theoretical accounts of any phenomenon, computational models of the described cognitive processes have been created to compare their performance against experimental data.

This line of reasoning is not yet been applied to AGL research on non-human animals, and most studies do not sufficiently take into account the problem that learning to distinguish between stimuli produced by different grammars, does not necessarily prove that the animal is capable of grammatical inference by using grammatical rules[4]. There is a growing body of research trying to elucidate the cognitive capacities of animals through the investigation of what grammars animals can handle (*if),* but these studies rarely control for alternative mechanisms resulting in similar behavior (*what* and *how)*. The response behavior seen in animals during AGL tasks may be a consequence of biases that the experimental design does not account for. It might be that an animal is grammatically incompetent yet possesses the ability to process acoustic similarities of the presented stimuli and make its apparent grammaticality judgment on this similarity [4].

1.2. Research questions

There is a large body of research concluding that various species of animals have the ability of grammatical inference and grammaticality. However, experimental designs often lack controls to account for other mechanisms than grammaticality. The lack of controls make the results reported in these studies debatable. However, their behavioral data can be used for explicit testing of alternative mechanism and models of how stimulus sequences are processed. To deepen our understanding and knowledge of the grammatical capacities required for animal grammar learning the following research-questions lay the foundation for this thesis.

Research question 1: Human artificial grammar learning researchers acknowledge that being able to distinguish between grammar-produced stimuli, does not imply grammatical competence, as it might not be the underlying grammatical rules that are being learned. *What effect does the assumption of grammatical competence, in the form of rule-learning, in animal AGL experiments have on the conclusions these studies draw?*

Research question 2: So far there is little evidence that there is a qualitative difference between fundamental neural mechanisms in animals and humans. Therefore, it may be that theoretical descriptions of the mechanisms thought to explain artificial grammar learning in humans are also

applicable to animals. *Can computational models of theoretical accounts of human AGL, other than rule-learning, account for the performance reported in animal AGL studies?*

Goals: Answering research question one requires an analysis of the animal artificial grammar learning literature. An investigation into *how* research is conducted, *why* it is conducted and *what* it concludes is done to see to what extent the field has rooted assumptions of grammatical inference and lack controls for other plausible mechanisms. Following that investigation, research question two will be addressed by creating computational models based on theoretical accounts of human AGL performance and by training those models on animal behavioral data. It is important to note that due to the novelty and exploratory nature of this project, multiple computational models have been explored yet only one has been implemented and been subject to experimentation. Hence, the research conducted here should also be seen as a primer for further research on modeling the cognitive processes underlying animal AGL.

1.3. Outline

This chapter has served the purpose of introducing the reader to the motivation for this thesis, as well as posing two research questions that will be attempted answered in this thesis.

Chapter two serves as an introduction to the AGL paradigm by reviewing the literature on human and non-human AGL research. The goal of this chapter is to answer research question one through a clarification of inconsistent use of terminology in the literature; a historical presentation of the AGL paradigm; exploration of the theoretical accounts of AGL; and an investigation into animal experiments that use the AGL paradigm, with a focus on the problem of perceived grammatical inference.

Chapter three's primary goal is to lay the foundations for addressing research questions two. First, a general introduction to computational and cognitive-modeling is given. It discusses what models can say about real-world phenomena; why modeling is necessary; and how one can utilize computational models in science. Then the relevance cognitive-modeling has to AGL and how modeling can be applied in this field is presented: elaborating on the same concepts presented in the first part of the chapter, but specific to AGL. Here are also presented a number of models based on the theoretical accounts of human AGL. The emphasis of this chapter is on giving the reader an understanding of the relationship between cognitive modeling and artificial grammar learning.

Chapter four is a presentation of the methods used for experimentation in this project. This chapter encompasses a description of the data, a presentation of a method for grammatical induction, and a number of in silico replications of animal AGL experiments together with their results and discussions.

Chapter five presents the results from the experiments in chapter four, in a more general form.

Chapter six discusses the strengths and weaknesses of the experiments and the models used, as well as a more general discussion on animal AGL research.

Finally, chapter seven addresses the two research questions of the projects and evaluates them in a concluding statement. Here is also noted the potential contributions this research has given to the field of AGL and outline possible directions for future research.

2 A review of Artificial Grammar Learning

To be able to understand the purpose and background of this research, it is important to consider the history of the artificial grammar learning paradigm: where it has come from; what its intentions were; its many adaptions; the way in which results have been interpreted; and how this holds up to contemporary research and methodologies. Therefore, the first aim of this chapter is to give a thorough presentation of the artificial grammar learning paradigm and the research in which it has been applied. To reach this aim, the reader is first introduced to some common terminology before the historical background of the paradigm is discussed. Following this, an exploration of theoretical accounts of artificial grammar performance in humans. Thereafter a critical elaboration on *how* animal AGL research is conducted, *why* it is conducted and *what* it concludes is given before it is finally argued that computational modeling is a natural next step in the groundwork for AGL research.

2.1. Terminology

The artificial grammar learning paradigm relies on a set of terms that are commonly used in everyday language. The everyday-language definitions deviate from the definitions of the paradigm, and there is also inconsistent use of terminology in the literature. Therefore, to avoid confusion regarding terminology and to keep terminological-consistency throughout this thesis, an introduction to the paradigms terms and jargon seems natural. If the reader is familiar with the AGL paradigm, this section can be skipped.

In everyday language, a *symbol* is usually regarded as a representation of something else. In the AGL paradigm, however, a symbol is a single or a set of meaningless letters, figures, characters or marks. Symbols are the most primitive unit of a language. They are the basic building blocks of the language being investigated. If one is investigating words, then the letters of the words are the symbols, or if one is investigating sentences, then the words themselves are the symbols. A sequence of symbols is called a *sentence* or a *string*. In the literature, the symbols of choice are usually dependent on how the stimuli are presented to the participant. Auditory strings have mostly symbols representing syllables [13]–[15], visual symbols include abstract tiles [16], [17] or letter shapes [8], and some studies are using tactile stimuli such as vibrotactile pulses as symbols [18]. In this thesis, the words symbols and primitives are used unambiguously.

The terms *'grammar'*, '*language', 'artificial grammar*, '*artificial language*, '*finite state language'*, and '*finite state grammar'* are in the literature used as meaning the same thing. Even though the only distinction that will be made in this thesis is the differentiation between 'grammar' and 'language', it is important to note that an artificial language does not imply it is a finite state language. A finite state language has its roots in formal language theory, while an artificial language does not necessarily adhere to any logic or mathematics and could, therefore, be anything. A *finite state language* is defined by Chomsky and Miller as the full set of strings created from a finite state grammar[19]. A

finite state grammar is a finite a set of rules that acts as a string generator. A grammatical rule in a finite state grammar has three specifications. First, the description of the current state of the string; second, the to-be-generated symbol; and third, the description of the string after the rule has been applied[19]. The terms *production rules*[20], *finite state generator*[8], *'generative grammar'* or simply *'grammar'* are often begin used in the literature with the same meaning as a finite state grammar. In this thesis *grammar* and *language* are used interchangeably with *finite state grammar* and *finite state language*, unless specified otherwise.

Grammatical inference or *grammar induction* refers to the process of learning the grammar underlying a language. *A grammaticality judgment* is a judgment on the well-formedness of a string, based on a rule-based representation of the grammar producing the string.[21] A string is well-formed if it is part of a language as described by a finite state grammar, and ill-formed if not. Here is a list of commonly used terms meaning the same as well-formed and ill-formed, in their respective pairs: (Grammar-consistent, Grammar-violating)[4], (syntax-conforming, syntax-violating)[22], (consistent, violating)[23],(grammatical, ungrammatical) [24], and (G, NG) [12]. If a human, an animal or a machine is capable of making correct grammaticality judgments, this entity is said to have *grammatical competence* or to be *grammatically competent.*

2.2. The history of artificial grammar learning

This section gives an introduction to the historical context of the artificial grammar learning paradigm, and what it has been utilized for since its conception in the late 1950s[8]. The reason for presenting the paradigm in such a historical way is for the reader to understand how the paradigm has evolved into what it is today and why these evolutionary steps have occurred.

What is a rule?

Even though one cannot always explain how we do it, humans have an intuitive feeling for how to discriminate between two objects, concepts or anything else. To do so, it is speculated that we have representations of abstracted rules in our brains that performs a form of similarity matching between the two things in question. Say for instance that you are tasked with distinguishing a bird from an insect. To do such a task, you put down simple rules that have a high chance of distinguishing the creature. Rules take the form of questions and, in the case of distinguishing birds from insects, essential questions would be something like 'does it have a beak?' or 'does it have feathers?' Such categorization rules are therefore simple mental operations that allow one to categorize an object by looking only at parts of it. We have specific categorization rules or questions for almost all the decisions we take and thoughts we make: To answer the question "is this number odd?", we only look at the final number in the sequence, not the number as a whole; To answer if a creature is a bird or an insect, we might only check if the creature has feathers, or if it has a beak. Some rules are more important than others, and some constitute what is called *critical feature* rules. A critical feature rule is

more resistant to change than other rules. Say for instance that some catastrophic event transformed the physical features of a bird into looking like an insect, but the event has left the bird still capable of mating with other birds. How do the questions asked previously hold up in this scenario? It has no beak nor any feathers, yet still, most would classify it as a bird. Therefore, that birds should be capable of mating with other birds, makes mating a critical feature of the bird category[25].

Project Grammarama

Throughout the 1950s and 1960s, American psychologist George Miller was interested in how humans induce such rules for categorization through observation and without explicitly being told of them. To explore this phenomenon, he launched 'Project Grammarama' at Harvard University. Project Grammarama was the first of its kinds in its attempt to understand how it came to be that participants who were exposed to some stimuli created from a finite state grammar, were able to recognize and distinguish between novel stimuli consistent with the finite state grammar and novel stimuli that were inconsistent with it [26]. In simpler terms, he wanted to investigate how humans intuitively learn the grammatical rules created from a finite state grammar.

In Miller's study that sparked project Grammarama, the participants in an experiment were told to memorize a set of presented strings and then a short time later attempt to recall as many of the presented strings as they could. The participants were not told how the strings had been created. After this familiarization phase, the participants were asked if they considered novel strings following the same grammatical rules as the training strings were grammatical or not. The results of the experiments showed two things. Firstly, participants were able to recall strings that followed the same production grammar at a higher frequency than recalling strings that were generated from a random grammar. And secondly, participants were able to correctly discriminate between previously unseen strings that were consistent with the grammar and strings that were violating the production rules of the grammar. From these results, Miller concluded that the participants encoded memory segments based on the common characteristics between the strings and that these memory segments were used to calculate a probability of what letters of the string would most likely occur together[8]. Miller considered these memory fragments to be a representation of the rules of the grammar.

The experiments conducted, and the results produced by Project Grammarama were both innovative and clever. However, the body of research-questions the program was addressing was too immense for a single lab to handle. The vastness of the questions being addressed was partly due to Millers own visions and goals: that if one is to study language acquisition in humans, it would be much more profitable to study the use of semantically based languages, and not languages limited to only syntax[27]. Even though the lab discontinued work on artificial languages and grammars, the project should in hindsight be viewed as a grand success as it is a significant contributor to the creation of the fields of psycholinguistics and cognitive science.

Reber's doubts

In contrast to Miller's emphasis on the importance of semantics in the study of language acquisition, the American cognitive psycholinguist Arthur S. Reber saw a use for non-semantical based languages. Reber assumed that the capability of classifying strings as grammar-consistent or grammar-violating would not be dependent on an explicit process during decision-making. His ideas dispute Miller's explanation that the participants were encoding the similarity of characteristics between the strings as segments in memory. Reber argued that if the participants are indeed capable of encoding the rules or string-similarities into memory, then they should also be capable of explicitly expressing these rules. This, however, proves not to be the case[9]. Reber replicated Miller's experiments with a few adjustments. Whereas in Millers experiments participants were told during the memorization phase that there was a structure to the presented strings, Reber did not. In Reber's experiments, they were told of an underlying structure during the testing phase only. This means that in Miller's experiments, the participants had the explicit goal of trying to learn and remember the underlying structure, while in Reber's experiments they were merely trying to memorize the strings. During the test phase of Reber's experiment, participants were shown previously unseen strings that were consistent with the grammar and strings that were in violation with the grammar and asked to put the string in either of the two categories. It turns out that the participants in Reber's experiment can make a correct classification between grammatical and un-grammatical strings with a frequency well above chance. In addition, participants are unable to express the strategies they used when determining the grammatical correctness of a string. These results confirmed Reber's assumption that one is not dependent on any explicit process to pick up on the structural regularities of an artificial grammar, but that these processes are implicit in their nature[9].

Reber's initial paper describing the modified experiment of Miller [9] presented the topic of implicit learning as a field of study on its own, using the artificial grammar learning paradigm as its tool of measurement. There is a large body of research conducted that agrees with Reber's conclusions of the implicit nature of the learning that takes place in an artificial grammar learning task [11], [28]–[30]. And the results of Reber's original work led to several interesting research questions. In a paper following Reber's initial publication on implicit learning, he took note of two questions of interest[31]. Firstly, *what* are the participants learning when they are presented with the grammar produced stimuli. And secondly, *how* is the participant learning what that is being learned. These questions will be focused on in the next section.

2.3. Theoretical accounts of artificial grammar learning

This subchapter outlines some of the main theoretical accounts of artificial grammar learning, with regards to *what* participants are learning and *how* they are learning it. While there are three main categories forming these theoretical accounts: rules, chunks, and similarity, the focus here is on the

contrast between rules and chunks. This is done due to the literature on similarity being highly ambiguous and provides little value to the remainder of this thesis.

2.3.1. The *What*: Rules or chunks?

There are a number of suggested theories as to what participants are learning in an artificial grammar learning task. Reber proposed that the participants in his experiments were learning and representing the underlying structure of the presented strings using rules[9]. And that the knowledge of the rules can be inferred by testing the participants' performance in a task to discriminate between grammatical and ungrammatical strings. However, a problem that arise from taking a rule-based approach to the *what* of AGL is that one has to account for the possible grammars the participants' might infer from the presented strings. As previously stated, the entire idea of the AGL paradigm is that the grammar and its full language is known to the experimenter and that the participant is the only unknown variable. The rule-based approach, however, falls in on its own argumentation in that participants might be perfectly able to extract the grammar from the strings and store them in the form of rules. But the grammar that they do extract would naturally be different from the experimenter's grammar because they have not been presented with all the possible strings. And therefore their representation might have 'illegal' rules compared to the original grammar. Another, and maybe even more important aspect of the rule-learning approach, is what these rules consist of. Johnstone and Shanks [32] proposed that the rules might be a representation of the transitional strength between symbols. A string would, in this case, be classified as grammatical if the symbols in the string have a high transitional strength. Dulany et al[33] argues that participants learn their own 'personal set of conscious' rules which help them make their grammaticality judgment. A personal set of conscious rules is meant to be smaller, less consistent rules than for example transitional strength, such as "does this string start with a certain character" or "does the string contain this specific bigram" and so forth. Whatever the representation of the rules are, it becomes bold and unjust to infer the learning of rules based on the successful discrimination of a set of strings, and the general experimental design to investigate rule-knowledge needs to be improved[12].

As previously discussed, Reber's idea that participants are implicitly learning rules contrast in some ways with Miller's explanation of the phenomenon. Miller argued that the participants in his studies were learning to identify similar features in the strings, and encoding these feature into memory. It can be argued that his explanation could be interpreted as identifying the rules of the grammar and then encoding these rules into memory. However, others have gone further and argued that it is fragments or chunks of the stimulus that are being encoded into memory[10], [11], [34]. Servan-Schreiber[35], [36] introduced the competitive chunking hypothesis which nicely shows what is being learned in a chunking approach to AGL. As the participants are being exposed to the training stimulus, they pick up on co-occurring symbols in the form of bigrams. These bigrams are what makes up the chunks or fragments that are stored in memory. Then bigrams co-occurring with symbols or other bigrams are

being recognized, forming trigrams, four-grams and eventually the full strings of the grammar. An observation not accounted for by the chunking approach is that participants exposed to grammar produced stimuli are capable of transferring their knowledge to a new set of strings made from the same grammar but with different symbols. It becomes difficult to imagine how a chunking hypothesis could explain the transfer of grammatical structure without implementing rules describing the grammar[10].

2.3.2. The *How*: The algorithms for extraction

Given the extensive amount of research that has been done with regards to rule-knowledge, it is surprising how little the literature discusses *how* the rules are extracted or learned from the training stimulus. Johnstone and Shanks[32] proposal of transitional strengths points towards extraction of rules through some form of statistical learning that calculates the distributional statistics of the stimulus. But other studies simply refer to the mechanism of 'implicit-' or 'incidental-' learning to describe the phenomenon, without any consideration to the algorithms that make up this learning. There have been attempts to simulate the performance of the 'implicit' learning using serial neural networks[37], [38], with a fair amount of success. These studies attempting to model AGL performance, claim that because of serial recurrent neural networks sensitivity to statistical regularities, implicit learning is, in fact, statistical[39]. In contrast, the *what* of the chunking and fragmentation proposal cannot be explained without accounting for *how* the chunks are being extracted. The *how* of chunking and fragmentation is explained as progressively building chunks from smaller chunks, resulting in chunks that represent whole strings from the presented stimulus.

The *what* and *how* of AGL will continue to be a topic of investigation for a long time. However, Perruchet and Pacton[39] claim that even the advocates of statistical approaches are beginning to acknowledge the idea of chunking, or at least that there is some form of psychological unit being created during learning. It could be that statistical computations are responsible for the formation of abstract rules, or the computations could even play a role in the formation of chunks and their strengths. Perhaps chunking does rely on computational statistics, or maybe chunking is merely the result of fundamental processes that add up to results similar to that of statistical computations. This will be discussed in chapter three.

2.4. Artificial grammar learning in animal experiments

As shown in the previous sections, artificial grammar learning was initially introduced as an experimental paradigm to investigate the phenomenon Reber referred to as *implicit learning*[9]. Since then, the paradigm has been used to examine a broad selection of possible learning mechanisms, both explicit and implicit and in a vast number of species and several scientific domains. In the animal literature, most studies assume that the animals are learning the production rules of the grammar, or abstracted versions of these rules. As shown in the previous subchapter, this is a deviation from the

human AGL literature, which is acknowledging other mechanisms resulting in the same grammaticality performance. It can be argued that the use of AGL in animal experiments has not yet matured, and is, therefore, investigating not so much the *what* or the *how*, but rather *if* the animals are capable of learning such grammars. This section is therefore dedicated to showing how, why and for what the AGL paradigm has been used in animal research. It is focused mostly on the conclusions these studies draw and the significances they claim. To build a truthful argument for the significance of this thesis, and to not mislead the reader, it becomes important to inform that the intention of this subchapter is not to dismiss or ridicule the research that has been conducted. The intention is to lay the foundation for an argument as to why the animal AGL paradigm needs to adopt the ideas presented in the human AGL paradigm. This is done by illuminating methodological errors, questionable assumptions and overstated significances found in animal AGL research.

2.4.1. A popular contemporary paradigm

There is a number of reasons why the artificial grammar learning paradigm has grown to be such a defining paradigm in the field of cognitive science. Firstly, the experimental procedure and design are both simple and intuitive. The overarching procedural structure of an AGL experiment consists only of a training phase and a test phase. During the training phase, the participants are asked to memorize a set of strings which are, unbeknownst to them, generated from a finite state grammar. After the training phase has ended, the participants are informed of the underlying grammatical structure of the strings and asked to make a judgment on the grammaticality of a previously unseen set of strings. This set of strings contains strings that are consistent with the grammar that produced the training strings, and strings that are either randomly created, has violations on specific transitions, or is made from a different grammar. The participants' performance is then evaluated based on the number of correctly classified strings. Though the participants can usually not explicitly report on the strategy they used for making the grammaticality judgment, their performance is well above chance[40]. The genius of the paradigm lies in the experimenter's complete knowledge of the finite state language – meaning all of the strings a grammar can produce – and of the strings which they train and test the participants on. This means that the participant is the only unknown variable in the experiment, which logically this yields that an experimenter can present a specific subset of string to the participant to investigate their ability to infer, represent and use the grammar.

Secondly, AGL blends well with most fields' methodological approaches. The AGL paradigm is mostly defined by its abstract methodological procedures, and not by the content of the grammar, the strings nor the symbols. This implies that the paradigm is not limited to the investigation of the cognitive processes related to grammar and language, but yields itself useful in a wide variety of experiments and research addressing questions related to sequence learning and pattern recognition. It is, therefore, maybe surprisingly due to its intention to investigate human implicit learning, also very well suited to investigate the mechanisms and cognitive processes that underlie pattern learning in

animals. Utilizing AGL in experiments on non-human animals yields insight into the evolutionarily conserved processes between various species[6], and allows for interesting research into the relationship between animal vocalization and human linguistics. The paradigm also benefits from using low-technology techniques, such as operant conditioning [41] in animals or card-displaying as in Millers original experiment [8], and high-technology techniques such as fMRI[42] or EEG [23].

2.4.2. Animal vocalization and the uniqueness of human language

Human language stands out in numerous ways on the feature-list of animal vocalizations. Semantics and the complexity of the language we learn are unique features contributing to the list. The question of how these features, both linguistic and biological, have evolved remain, arguably, an open question. The interplay between the linguistic evolution and the biological evolution becomes not only a source for the complexity of the human language faculty but also makes these features rather challenging to investigate. The implications of not having a fossil record in the examination of the human language faculty's evolution, make the evolutionary narrative on the topic not much more than 'fanciful storytelling'[1], [43]. Bypassing the difficulties of missing fossil records is, however, possible through the mapping of conserved evolutionary features across species. Comparative work remains the most compelling method for investigating the limitations, ranges, and features of the human language faculty[43]. Whether one is a supporter of the proposition that language is a bi-product of features evolved for internal information processing, or a proponent for the view that language is 'just' an effect of a species general-computational mechanisms and that these mechanisms and abilities in human are more advanced: One uses non-human animal research to argues one's case through comparative work[44]. Arguably all of the sides in the uniqueness of human language debate agree that the evolution of language, as we know it, must have evolved from the already existing computational abilities of humans. This implies that it is the investigation of the underlying and more fundamental mechanisms of the computational abilities of humans and non-human animals that will be one of the more substantial contributors to the advancement of the debate surrounding both the complexity and uniqueness of the human language faculty.

2.4.3. The many answers of experimental AGL animal research

There is a large number of studies which utilize the AGL paradigm to address the mapping of non-human animals' computational abilities from an experimental point of view. Most animal experiments are investigating animals potential to learn rules. Murphy et al[45] trained rats on a set of rule-consistent sequences of stimuli through a reinforcement task, and reports that the rats are capable of distinguishing between rule-consistent stimulus (XYX) and stimulus that does not obey the rule (XXY, YXX). They also report that the rats a capable of transferring the learned rules to novel stimuli. Toro and Trobalón also report on rule learning in rodents, but in this case more specifically on their ability to process statistical computations[46]. Reportedly the rats in this study were trained to press a

lever at the presence of a speech stream before they were tested on how part-words, non-words, and words affected the rats' lever-pressing performance. Supposedly, the rats' response behavior was sensitive to the frequency of co-occurring syllables, but not to the transitional probabilities in the speech. The Murphy et al. paper[45] claims that their study shows that rats are capable of transferring the grammatical rules to new novel stimuli, and therefore, like humans, they can transfer structural knowledge from sequential experiences. Claiming such a significance implicitly states that there should be some evolutionarily conserved feature between rats and humans that can account for such a good analogical performance. Such findings would have major implications if one could exclude the possibility that the rats' responses have been determined by factors not related to the rules at all. Corballis [47] criticizes Murphys et al.'s claims and argue that there is little resemblance between what the rats have learned in this experiment and what humans are doing when using language. He argues that the rats' discrimination performance could be explained by their ability to match unique pairs in the stimuli. Meaning that the rats could be able to distinguish between grammatical and ungrammatical strings based on only the grammatical strings (XYX) having the same first and last letter, which is different from the ungrammatical strings (XXY or YYX) which has matching pairs. This shows that such discrimination need not be dependent on any such rule-learning which is claimed in the paper. Toro and Trobalón's paper also create their own complications when they address the fundamental mechanisms underlying language use and acquisition. Their trouble lies in the lack of a connection between being sensitive to statistical computations and any evidence for a conserved evolutionary component between rats and humans. The authors do, however, point to Hauser et al. [48] and suggest that mapping the computational abilities of animals through the investigation of their sensitivity to statistical regularities, may give answers on some animals limitations for developing linguistic abilities. While this surely is a solid argument, such a capacity mapping does not contribute to the questions of what the underlying mechanisms for language are, but merely what these mechanisms limitations are. And in this way argue against the conclusions they draw themselves.

Focusing on an entirely different group of animals, birds, ten Cate and Spierings[5] compared zebra finches and budgerigars abilities to abstract grammatical-rules from one string-set and used the rules on a novel string-set. In the paper, the authors' report that the budgerigars can generalize their rule learning from one string-set onto another set of strings and that this is an indication that the budgerigars are capable of a level of abstraction comparable to human analogical reasoning. To note, the zebra finches were not capable of making this rule abstraction but appeared only sensitive to novel strings containing symbols from the training set at their ordinal position. Ten Cate and Spierings suggest that the ability of analogical reasoning, as supposedly seen in the budgerigars, is an ability the bird has in common with human infants and that this ability has long thought to be unique to humans. Another study which reports on complex mechanisms previously thought to be unique to humans is a 2006 study by Gentner et al. [7]. In this study, European Starlings are trained to recognize strings

generated from a recursive, self-embedding-context-free grammar. Recursion, which is a property that has been suggested to be one of the distinguishing features between human language and animal vocalization[48], is represented in the grammar when a symbol or a sequence of symbols can call upon a repeated production of itself. The starlings who were trained using a go/no-go operant conditioning procedure were able to distinguish acoustic patterns from a context-free grammar (A^2B^2) and a finite state grammar ($(AB)^2$). The authors claim that this study shows that there is at least one non-human animal capable of learning a self-embedded recursive structure. There are a number of critiques directed at the Gentner et al.'s study, and the paper's conclusions. Firstly, the strategy the starlings used for distinguishing the two grammar was not sufficiently controlled for. The strategy for discrimination between the two grammars can be explained by much simpler mechanisms than recursion. Corballis[49] shows that such a strategy could be as simple as requiring no more than a counting mechanism and good enough memory to remember the number of elements. Suhara and Sakurai[50] show through a Bayesian classifier that the starlings need not have any understanding of recursion to account for the reported performance. Both the counting mechanisms and the Bayesian classifier shows through examples that simpler mechanisms can account for the starlings' response behavior. Van Heijningen et al[51] replicated the Gentner et al.'s study using zebra finches instead of starlings. They found that zebra finches were also capable of learning the same discriminations as the starlings. However, the authors argue that to be able to show that the birds have learned to discriminate based on the recursive structure of the stimuli, they would also have to be tested on strings using the same grammars but with novel symbols. This proved too difficult for the zebra finches to learn, which van Heijningen et al. argue is evidence that the discrimination is based on phonetic cues, and not on syntactic/grammar generalization. This, in turn, empowers the skepticism of Gentner et al. 's original conclusions, and the assumptions of rule-learning.

Taken together it seems clear that some conclusions on grammaticality in animals are debatable at best. In many cases, more straightforward explanations, often based on non-grammatical approaches, have not been sufficiently controlled for. Some animal studies are conducting their AGL experiments using a habituation phase, where they familiarize the animals with a string set put together by strings made from a finite-state grammar. Either after a certain amount of time or some other criteria have been met, the animals are tested on their ability to classify grammatical strings from un-grammatical strings. If the animals show behavior, or other measurements indicate that they are able to differentiate between the grammatical and ungrammatical strings, the researchers leap to the conclusion that the animal has learned the rules of the grammar. This conclusion, or claim, is however questionable at best. Beckers et al[4] point out the importance of differentiating between the claim of learned grammatical rules from a string set, and the recognition and familiarization of acoustic features in the same set. They identify a number of studies where strategies based on lower-level sensory features are ignored but could act as an equally good explanation for the observed behavior during the test phase,

as the explanation claiming learned grammaticality. While Abe and Watanabe's report[52] showing that Bengalese finches are capable of discriminating auditory stimuli based on syntactical rules might be correct, one should, as Beckers et al. [22] demonstrate, be wary about making such claims. Higher order claims, like the ones Abe and Watanabe make, do have a rich potential in the scientific community if they turn out to be correct but yield little value when there are simpler hypotheses, in this case: acoustic similarity matching, that could explain the observed behavior as well. Many, if not most, animal AGL studies could be explained by either a simpler hypothesis or by a more fundamental strategy. Contrasting the argument for simpler mechanisms, ten Cate [44] argues that even though the emergence of learned grammaticality can be explained by these simpler mechanisms: It does not imply that the animals are unable to use and identify higher-order regularities in strings. While such an argument undoubtedly has some truth to it, one can only consider the identification and confirmation of higher-order learning abilities an increasingly difficult task due to the implicit nature of knowledge in these experiments. If the statement that higher-order regularity identification might be present even though simpler and more fundamental explanations exist holds true. Then it would be implicated that the human AGL would treat their performances in the same manner. However, the investigation into non-grammatical explanations is slowly and steadily finding its place in human research and experiments. The human AGL research community is accepting that there are underlying low-level mechanisms that give rise to the perceived grammatical rule inference in experiments. A more theoretical and bottom-up approach to investigate and explain the phenomena of artificial grammar learning has begun. Where animal AGL seek to explain behavior using higher-order mechanisms, the human AGL world pursues lower-order mechanisms, and this is where the main deviation between human and non-human animal AGL research lies.

2.5. Summary and the next step

The artificial grammar learning paradigm has over the last fifty years given rise to an extensive body of research across a diverse set of disciplines, and across a vast set of species. Researchers who use the AGL paradigm in animal experiments often claim a significance about the fundamental questions in the uniqueness of human language debate. However, such experiments are often not optimally designed to unequivocally answer the questions which they are addressing[22]. On top of this, they tend to claim that animals are capable of learning grammar or the production rules of the grammar, whereas such claims seem to be part of a skewed explanatory narrative as there lies evidence for simpler explanatory mechanisms. The effects of having an underlying assumption of rule-learning in the animal AGL paradigm results in a lack of controls for other underlying mechanisms that are just as, if not more, likely to explain the animals' performance. The problem the animal AGL paradigm faces is its current attempt to explain cognitive capabilities, such as grammatical capabilities, without looking at the more fundamental mechanisms that the animal possesses. Non-grammatical explanations based on low-level sensory features and other information processing principles are more

or equally likely to account for the phenomena as higher-order grammaticality explanation. Further investigation and exploration into the cognitive capacities for grammar in non-human animals will strongly benefit from researching the fundamental mechanisms that can account for the learned or emergence of learned grammar. By modeling low-level sensory features such as chunking and fragmentation, or information processing principles such as memory and transformation one can get a deeper insight into how the higher-order subsystems of grammar-learning behave; as well as assisting in the creation of new hypotheses and experiments through models predictive nature. The cognitive modeling of computational abilities of animals, therefore, becomes the next natural step in the exploration of evolutionarily conserved mechanisms between humans and animals.

3 Computational and Cognitive Modeling

Laid out in the previous chapter is what the AGL paradigm is, how it has been utilized in the human and animal world and possible theoretical accounts of the paradigm. This chapter is dedicated to the exploration of cognitive mechanisms, in the form of computational models, which are hypothesized to account for grammar learning in humans. To begin this exploration, it is important to first reflect on what computational and cognitive models are and what use they have for science and research. Therefore, the first section of this chapter gives a general reflection on what computational modeling is and addresses the discussion on what value cognitive models have in explaining real-world phenomenon. Following this discussion, the scope is narrowed to focus on models created to account for AGL performance in humans, with a focus on models for chunking or word segmentation.

3.1. What are computational models?

Science, it can be argued, is a systematic way in which humans organize and build our knowledge in an attempt to make predictions on nature. This collective knowledge is made up of theories and frameworks made to understand the world on a more fundamental level. We create descriptions and information systems from our reasoning about the world and improve upon these descriptions through further exploration and experimentation. One way to improve our exploration of natural phenomena is through computational models. Computational models are mathematical specifications, or algorithms, representing our understanding of some phenomena. Computational or mathematical models give scientist the advantage of being able to observe variations more efficiently through the use of computer simulations. To gain a better understanding of the concept of computational modeling and how it works, let's consider how a chemist is trying to figure out how the various chemicals of a certain concoction contribute to the final result. Our chemist reasons that a solution to figuring out their interactions would be to make the concoction ten times and leave out a single chemical each time. Doing so would, however, be excessively time-consuming. A better alternative for the chemist would be to enter the ten chemicals in a computer model that explain what each chemical does and how it interacts with the other chemicals. After doing so, the chemist has the option to run multiple computer simulations, each with a different chemical being removed from the concoction. The chemist could quickly gain an overview of what the final concoction would look like if they removed two chemicals, or three chemicals, or changed the amount of each chemical. Computational models are powerful in the way they allow scientist and researchers to simulate a large number of variations of phenomena. In the situation describing the chemist, however, it is assumed that the interaction and details of the chemicals are mathematically well-described representation of their real-world counterparts. However, there are not many scientific fields that can describe their theories using well-formed mathematical formulas and should, therefore, be considered approximations of phenomena and not replications of them.

3.2. Analyzing cognitive models

One subset of models that naturally fall into the category of approximation models are models attempting to predict or comprehend cognition and the processes in which it encompasses. Models that approximate the cognitive processes in humans or animals are called *cognitive models*. Studying a process, cognitive or not, can be a rather peculiar task. The study of processes can take place at several levels of abstractions. A formal way of explaining the processes of picking up an apple, for example, could be to present a person with the task. And then measure the time and precision in which they perform the task, to estimate their behavior. Another behavioral prediction could be made by studying the pattern of firing neurons in a person performing the task. However, in isolation, neither of these would be able to fully explain how the process of picking up an apple works. The first example would only give an estimation as to what measures one could expect to observe, and the patterns of neurons firing in the latter example would not be directly translatable to the observed behavior. Even if it were possible to map out the entire brain of a person picking up the apple, it would be impossible to translate the firing of a single neuron to observable behavior. To cope for the many levels of abstractions in which one can describe a process, David Marr, a British neuroscientist, introduced three levels of analysis for understanding a processes[53]. The first, the *computational level*, describes the function that is to be computed, it describes whether the goal is to learn a function, make a prediction, estimate uncertainty, or to pick up an apple. The second level of analysis is the *algorithmic and representational level.* Here is determined the algorithms which are needed to reach the goal of the computational level, as well as the input and output to and from the algorithms. Lastly, there is the *implementation* level, which considers how the algorithms presented in the algorithmic and representational level can be physically implemented, is it in neurons in the brain or is it implemented in silicon in a computer chip? Each one of his three levels of analysis should be thought of as a fulfillment of the layer above/below it. The algorithmic and representational level should be describing how the computations in the computational level can be computed, and the implementation level should similarly describe the systems that implement the algorithms. Marr argued that these levels are essential when addressing questions regarding the behavior of complex systems. The three levels of analysis have been embraced by the field of cognitive science(s) and have become a popular paradigm in which to reason about complex systems.

Marr himself was an ardent critic of models that were not rooted in the computational level as an approach to understanding cognitive systems. Marr argued that models attempting to approximate cognitive processes should be rooted in questions addressing both the why and what of the computational level[53], [54]. The root of Marr's argumentation lay in the possibility of being able to mimic the performance of the system without having any understanding of the fundamental mechanisms underlying it[55]. It can be argued that Marr took a position on this topic before there was a clear distinction between cognitive models and artificial intelligence models. In the late 1970s the

rise of computational power for scientific use exponentially rose, and with it came theoretical frameworks to run on the computers. As Marr was a neuroscientist interested in vision, his goal was to make a coherent theory of vision, which meant creating models for explaining and exploring the cognitive processes that make up vision in humans. His goal was not to create the best possible model for vision but to replicate the model of vision that humans share. Contrasting this idea, others began creating models for vision that did not use theories of cognitive processes as their fundament, but used theories of computer science and mathematical optimization instead, while still calling it 'theories of vision'. It is difficult to believe that there are researchers today that would disagree in Marr's complaint that it becomes unreasonable to call models that fail to answer questions about the processes underlying vision for 'theory of vision', even if the models have the ability exhibiting the same behavior as the human visual system[55].

3.3. Statistical learning or cognitive processes

Marr's complaint regarding the theory of vision touches upon an issue that is found in all of science: given some reported behavior or observed phenomena, what is the nature of this behavior or phenomenon? It is an intrinsically difficult problem to deduce the nature of any phenomena, but cognitive mechanisms might be in the set of the more difficult phenomenon to investigate. To give an example of this, let's consider the computational abilities of humans. In 1996, Saffran, Aslin, and Newport introduced the idea that infants have the capability of computing statistical properties of speech patterns [15]. They showed this with an experiment testing if infants are able to separate words in a continuous speech stream using nothing but the statistical relationship between the constituent sounds. The infants can do so, even after just two minutes of exposure to the speech stream. Following this experiment, statistical learning has gained significant traction within the domain of language acquisition and is now arguably the dominant paradigm. Later that same year, an experiment showed that adults were capable of learning the words of a language where the word-boundaries are made solely from the transitional probabilities between two syllables[56]. This means that the participants that are capable of learning these words must have learned the transitional probabilities making up the distributional cues of the speech stream. And from this, it is suggested that the computation of transitional probabilities play an essential role in language acquisition. In addition to the two studies described above, other studies show evidence pointing towards infants, and adults, being sensitive to other statistical regularities in language, such as prosodic and phonological cues[57], [58]. However, being sensitive to transitional regularities does not mean that the mechanisms are statistical in nature. Saffran et al. [56] acknowledge the computational complexity arising from segmenting a speech stream without prosodic and phonological cues for word-boundaries, purely on a statistical basis. Still, they conclude that this process is in its nature statistical, without giving any consideration to how the statistical regularities are being extracted or stored. According to Perruchet and Vinter[59], such a conclusion is a victim of circular reasoning. Word knowledge, they argue, is simultaneously the pre-

requisite for detecting statistical regularities as well as being the consequence of this detection. Challenging this circularity and the computational complexity of statistical learning, Perruchet and Vinter introduced a computational model contrasting that of statistical learning. Their model, PARSER, has considerably less computational power and memory than what is estimated to be required for calculating the transitional probabilities between syllables. The model, which will be described in detail in the section *Models of AGL*, shows that it is possible to do word segmentation on the level described by Saffran et al[15] without the need for the computation of transitional probabilities or any other statistical regularity[59].

Borrowing from Marr's argumentation on models made for a 'theory of vision', one can argue that while statistical learning provides a good fit for the behavior observed during language acquisition, and the properties in which the participants are tested on are statistical in nature, the cognitive processes giving rise to behavior reflecting these regularities need not be statistical at all. To illustrate with a simpler example: imagine a small sensor sitting on a window sill. You are tasked with figuring out what the sensor is doing. By observing the sensor, which is all you can do, you notice that whenever the sun stops shining at it, a small light next to it switches off, and turns on again as the sun hits it. Therefore, you conclude that the sensory is reacting to the light from the sun. However, this is not the case. The sensor is detecting changes in temperature, and immediately as the sun stops shining on it, the temperature is lowered a tiny bit and the light switches off. While the sensor surely can be described as being sensitive to the light from the sun, and the light theory nicely describes the observed behavior, light has nothing to do with it with the nature of what it is detecting. The same goes for rule learning in artificial grammar learning tasks, and for statistical learning in language acquisition. If the interest is to explain the underlying mechanisms of any phenomenon through the observation of behavior, one needs to consider all levels of Marr's framework before jumping to conclusions.

3.4. The value of modeling

The goal of highlighting the difficulties in drawing conclusions from experiments on cognitive processes is to show that the explanatory value severely drops with the lack of reflection on what constitutes a cognitive process. In chapter 2.4 it was argued that it could be highly interesting that some birds are sensitive to the rules of grammar produced stimuli. But that making claims of grammatical competence from rules, should not go without experimental controls for other mechanisms or reflection on the entirety of the system in which grammatical competence is being inferred. This principle is generalizable to the investigation of all cognitive processes. An explanation that is merely interesting, or being able to fit the data well, does not give scientific justification for making causal explanations without proper controls and systematic evaluation and reflection. Modelers of cognitive processes are forced into these types of evaluations and reflections because of the explicit nature of models. Any assumption the modelers make are easily detectable as their

assumptions are explicitly described through mathematical formulas or computer code. This allows computational models to be subject to more rigorous testing and evaluation than what their purely theoretical or intuition based counterparts are. Taking a computational modeling approach to scientific inquiries can, therefore, offer a better intuitive account of a phenomenon. And a constant evaluation of such models based on new data can be used to offer an answer too which of several alternative accounts is the adequate one. Explicit computational models are for these reasons becoming more and more central in the investigation to unravel the underlying mechanism of cognition[60].

3.5. Models of artificial grammar learning

Where animal AGL research contrast with human AGL research, is in its determination to "identify the subsystems without giving detailed consideration to the mechanisms involved" [60]. Meaning that they consider animals capable of learning grammatical rules, without considering the underlying mechanisms. Human AGL researchers have begun acknowledging the intricacy of the system that is being investigated and is for this reason applying a more mechanistic, or bottom-up, approach to the problem, in addition to the traditional top-down approach. Though there is no theoretical consensus as to what mechanisms are responsible for the learning, or perceived learning, of grammatical-rules in humans, there exist a number of theoretical learning mechanisms claiming to account for AGL performance [12]. Evaluating all models that could potentially explain AGL performance is too broad for this project. Therefore, the scope is narrowed to models of word segmentation or chunk extraction. In the coming section, three models, or types of model, proposed as explanations for artificial grammar learning are presented along some of their applications and experimental results.

3.5.1. PARSER: A model for word segmentation

'PARSER: A model for word segmentation' is a computational model intended to show that the extraction of words from a continuous speech stream could be the result of memorizing chunks of the stimuli. And that these chunks then further guide the perception of new chunks[59]. While not explicitly stated, the PARSER model offers an explanation of word segmentation on the representation and algorithmic level in Marr's levels of analysis. The model has its focus on a memory unit containing a vocabulary of chunks together with the chunk's weight. The chunk weight is a representation of the chunks strength in memory. The model's algorithmic structure centers around one memory unit storing chunks and their weights, one action called perception, and two transformation processes, forgetting and interference.

The model takes as input a finite (theoretically it could be infinite) stream of symbols. The initial state of the model is the memory unit containing nothing but the symbols of the language. At each time step after the initial state, the model acts by processing parts of the input stream through the act of *perception*. When the model perceives, it selects an arbitrary value, X, that determines the size of the next perceived unit. The next perceived unit is then a combination of the X longest chunks in the

memory unit that has a weight above a predetermined threshold value. If the perceived unit is contained in the memory unit, then the weight is adjusted accordingly. Alternatively, the perceived unit gets added to the memory unit if it is not in there already. As a consequence of perceiving, the memory unit undergoes two transformation processes. First, all except for the currently perceived units, undergo a forgetting step in which all weights are decreased by a pre-set forgetting value. Then all of the weights of the chunks in which any of the substrings contained in the perceived unit are decreased. After the model has undergone these transformations, the cycle starts again, and in this manner the model uses its memory to shape the perception of the incoming stream and to progressively extract words from this stream. The output from the model is a vocabulary of chunks and their strength. How sensitive the model is to changes in any its six parameters (1, 2) the maximum and minimum length of a perceived unit, 3)the threshold for a word being,4) the forgetting weight, 5) the interference weight and 6) the initial primitive weights is not currently known due to the exhaustive task of mapping such a large parameter space[12], [61].

PARSER has been used to test a number of hypotheses. In the paper describing the model, Perruchet and Vinter[59] set out to challenge the conclusions drawn from Saffran et al.'s findings that adults can segment an artificial language having no prosodic cues or pauses for word boundaries[15], [62]. They replicated the experiments using computer simulations of the PARSER model and showed that the model is indeed able to replicate the results of the original experiment. Giroux and Rey [63] compared PARSER's performance to the performance of human participants in a language acquisition task. They tested adults on their ability to recognize chunks drawn from the exposure stimulus as valid after a two-minute exposure and a ten-minute exposure. Their results suggest that the chunking strategy PARSER use results in a good fit with the human performance data[61].

3.5.2. TRACX and TRACX2

TRACX[64], or 'truncated recursive autoassociator chunk extractor' is a connectionist approach to the problem of chunk extraction and sequence segmentation. The TRACX model focuses on what the authors call *implicit chunk recognition*, which means that it attempts to recognize previously encountered strings of symbols instead of predicting them, which is what most other connections models of chunk extraction do. TRACX is based on a neural network architecture called autoencoders. Autoencoders are used to gradually learn to compress input into a smaller representation of itself for then to decode this representation into its original form. The more often TRACX encounters a string, the better the encoding of the string becomes, and therefore produce a smaller error upon encountering frequent strings and larger error on infrequent strings. The implemented version of TRACX is a three-layered feedforward autoencoder network using a backpropagation algorithm as its gradient calculation for weight calibration. The network takes as input a stream of symbols, and initially 'perceives' two symbols from the stream. These two symbols are encoded in the network and are used for further perception of new symbols. If the error produced by the output nodes is lower than a set

threshold when a string of symbols is presented, the string is accepted as a frequent chunk, and the next input to the network continues where the accepted string ended. If the error is high, the string is rejected, and more of the input string is considered until an accepted string is found. In this way, the network learns first to encode smaller chunks of the input stream. As the error from the output decreases as a chunk is encountered more frequently, larger fragments or words containing the initial smaller chunks are gradually being recognized. The authors claim that the psychological and biological plausibility of autoencoders is well established, and therefore present TRACX as a biologically plausible model. However, the TRACX model relies on a symbolic if-else switch that uses a pre-set threshold to determine the rejection or acceptance of a string. For this reason, TRACX2 was created with the intention of improving the biological plausibility of the model[65]. TRACX2 follows the same architecture and procedures as the original TRACX model, but instead of relying on the if-else switch, it uses a weighted sum from the output layer to guide its acceptances or rejections.

TRACX and TRACX2 has been used to model a number of experiments. French and Mareschal[66] modeled a number of statistical learning experiments. In modeling Kirkham et al.'s [67] visual statistical learning experiment TRACX2 was trained on am input sequence that was shared the probability structure of the input used to train the infants. They showed that the model performs in similar ways as the newborns that were tested in the Kirkham et al. experiment, and that TRACX2 was able to accurately capture statistical learning in the newborns. In modeling an experiment by Slone and Johnson[68], TRACX2 was shown to be able to form illusory conjunctions. Such illusory conjunctions are commonly argued to be evidence for some form of statistical learning mechanisms. However, TRACX2 also captured a decreased accuracy on embedded chunk items, this feature is commonly attributed to mechanisms of chunking. Meaning that TRACX2 was able to produce results thought to be exclusive to chunking and exclusive to statistical learning within the same mechanism.

3.5.3. Simple recurrent networks

Simple recurrent networks, or an Elman network, is a type of neural network structure made for processing temporal or sequential data through a context layer[69], [70]. There are four types of units in an SRN: an input layer, a hidden layer, an output layer and a set of *context* units that are connected to the hidden layer. For every time step in the data, the context unit store the previous values from the hidden layer. This allows the network to 'remember' the previous step in the data. There is a fair amount of literature addressing the problem of word-segmentation using SRN, however, much, if not most, of the literature on the topic address the problem of word-segmentation as an optimization problem. Moreover, they remain agnostic as to whether SRNs could explain the cognitive mechanisms of word segmentation. Nevertheless, there are also implementations of SRNs that do addresses cognition and compare their results to those of humans[37], [71], [72]. Where the TRACX model learns by attempting to recognize the next chunk in the stream, SRNs learn by attempting to predict them and then attempting to correct its weights accordingly[64].

SRNs have been shown to fit nicely with human performances in a number of language acquisition tasks. Cleeremans and McClelland [37] used a serial reaction time task to investigate how an SRN obtained the grammar from finite state language and compared the results human performances. The human performance indicated that humans are limited to processing a maximum of four items, the SRN however, was eventually capable of correctly processing a higher number of items. SRNs have also been a somewhat popular model to compare to chunking models. In modeling how people learn the associations between adjacent symbols in strings, Boucher and Dienes[71] compared the results from an SRN, with results from a model made based on the competitive chunking hypothesis[36], and results from human experiments. They showed that the results chunking model provided a better fit for the human data than what the SRN did. The authors believed that the reason for why this is the case is because of how an SRN corrects itself. If an SRN is trained on a set of strings, its weights converge towards a representation of that string-set, therefore, if an already trained SRN is presented with a new string-set, it will automatically start the convergence towards weights representing this new string-set. Because of the limited memory of an SRN, it has to switch out the old representations if it wants to learn something new. The second finding from this study was that the SRN was more sensitive to and therefore learned to represent the forward probabilities more accurately. Another study by Perrcuhet and Peerman[73] confirms SRNs sensitivity to forward association, or at least compared to a chunking based model.

3.6. Summary and model selection

An evaluation on what constitutes behavior and what constitutes a mechanism shows that theorizing about complex processes, such as artificial grammar learning, requires an understanding of the process on multiple levels. Observed behavior does not alone provide evidence for the underlying mechanisms that drive this behavior. Manifesting an understanding of complex processes in cognitive and computational models have the potential to provide insight into the underlying mechanisms and processes that make up cognition. When it comes to animal AGL, the literature has a tacit assumption of rule-learning, however, there is evidence of chunking mechanisms in non-human animals[74]–[76], which serve as an equally likely explanation. Therefore, modeling offer an opportunity to investigate to what degree a chunking mechanism can explain results previously assumed to be the result of a rule-learning mechanisms in animals. The models presented in the previous section are all word-segmentation models made with the principles of chunking in mind. While all of them could prove themselves useful in modeling animal AGL, I have, for a number of reasons, chosen to use the PARSER model. First, PARSER is a well-known model in research on infants' language acquisition and has shown promising potential to explain their grammatical abilities, and is therefore appropriate to use from a comparative perspective. The second reason is transparency. PARSER is a symbolic model that gives explicit insight into every action it makes. That TRACX and SRNs are manifested in neural networks removes the transparency of the two, as neural networks inner workings and decision

making processes are difficult to interpret. The third, for why PARSER was chosen is that PARSER is a simple, intuitive and understandable model. It was designed with the intention of providing a simple explanatory mechanism to account for what was suggested to be learning from statistical distributions. The final reason for selecting PARSER is that its simplicity and transparency makes it reasonable to implement in a programming language of my choice. The implementation of the PARSER model is not being discussed in this thesis, but the implementation and the experiments conducted with it is publicly available in a GitHub repository[77].

4 Methods

Since PARSER is a model for word segmentation and does not give directly interpretable results for an AGL task, a method allowing PARSER to make grammaticality judgments on novel strings is presented. This process and the model is then used to replicate a series of AGL experiments. The data and experimental structure of each of the experiments will be explained as well.

4.1. A method for measuring the grammatical competence of a chunking model.

To evaluate a trained PARSER model, Perruchet and Vinter [59] introduced two performance criteria, a loose criterion, and a strict criterion. The *loose criterion* is met when all of the strings from the presented language are represented in the model with the highest weights, but also contains a set of non-words or part-words. The *strict criterion* is met when the *loos criterion* has been met, but all part-words and non-words have been filtered out. The PARSER model was built with the intention of showing that there are simpler and more straightforward (proposed) cognitive mechanisms that can produce the same results as a model using statistical learning. It is therefore strange to not see a reflection on their own argumentation when designing a performance criterion. There are two reasons as to why this is important. First, there seems to be a clear difference between the processes of extracting a string from memory and comparing it to another string, and that of making a judgment on a string's grammatical well-formedness. Second, a trained word segmentation model such as PARSER is nothing more than a vocabulary of strings and their weights. At best, the vocabulary contains only the strings which it has been trained on, leaving no room for generalization to novel strings. For these reasons, a new method for measuring the grammatical competence of a chunking model is introduced. Instead of asking if a string is in the vocabulary, this method allows experimenters to ask if a string is consistent with the grammar of the vocabulary.

4.1.1. Grammar inference as a cognitive mechanism

It is important to note that grammar induction is also a branch of machine learning algorithms attempting to learn the grammar of a formal language from a set of observations. However, these algorithms do not concern themselves with how it functions in the brain, but only with the most efficient way of learning the grammars. For this thesis, the cognitive aspect is essential, and it is, therefore, necessary to present the assumptions about cognition that is made. Another important remark is that this method should be concurrent with the way the PARSER does its segmentation.

One aspect of cognition that the presented method does share with more common machine learning techniques is the aspect of optimization. The model wants to optimize its knowledge of words, but also to optimize its generalization to novel words. The learning of a new chunk from two smaller chunks, should therefore not result in loss of the knowledge from the smaller chunks. From this, the first assumption I make is:

1) **Assumption 1:** There should be no loss of information when that information is used to learn something new.

Assuming that the first assumption is correct, it is implied that when learning a new chunk from two smaller chunks, the information about the smaller chunks should not be lost. Analyzing this implication from an optimization point of view, it can be argued that the model is attempting to optimize its memory, and therefore favors larger chunks over smaller as there is more information contained in a larger chunk. The information of the two smaller chunks is still contained within the larger chunks, and one has to have a process to extract it from the larger chunks. Therefore, the second assumption is:

2) **Assumption 2:** Memory is expensive, processing is cheap.

The third assumption to be made is needed more for the computational aspects for implementing assumption 1 and assumption 2, and is best explained and justified through an example. Imagine that one is to travel from point A to point D, and to do so you have to travel through points B and C. According to the assumptions above, and the mechanisms of chunking, one would not directly remember each single path: (A-B), (B-C), (C-D), one would remember (A-B-C-D). However, when one is tasked with the actual traveling, (A-B-C-D) would be split into the smaller sequences (A-B), (B-C), (C-D). Generalizing this principle to chunks and symbols, it is explicitly assumed that:

3) **Assumption 3:** A chunk is more than the chunk itself. A processed chunk is the combination of all the transitions between symbols in that chunk.

From these three assumptions, three processes that occur at the time of a grammaticality judgment are presented: extraction, matching, and building.

Extraction is the process of extracting all the chunk knowledge the vocabulary has. If one has the string ABC and process it to extract all knowledge contained in it, the result would be the following transitions: (A-B), (B-C), (AB-C), (A-BC). Assuming that processing is cheap, extraction is done on the entire vocabulary, as well as the input string.

Matching is the attempt to match the extracted chunks from the vocabulary to the extracted chunks from the test string. The result from matching is the set of transitions that are legal that are contained in both the extracted vocabulary set and extracted input set.

Building is the attempt to use the information about the matched sequences to build the input string. If all extracted transitions from the input string can be found in the extraction set, the string is to be considered grammatical; if there are however missing transitions, the string is considered ungrammatical.

4.2. Datasets

Beckers et al[4] wrote a paper critiquing stimulus design in AGL studies on animals. The paper reviews the stimulus from eleven different AGL studies. The data that was used in Beckers et al.'s paper is available online and was collected for use in this project. The data is a collection of string sets and their respective string-categories. In this project, the strings, the categories and a variable called *readingframe* have been used. Other variables such as *labelcolor*, *isduration*, and *tokenduration* have not been used.

4.3. Experiments

The goal of this thesis is to test if a simple learning mechanism such as chunking can account for AGL performance in a wide variety of published animal experiments. If so, the animals could be using memory fragments and processing to make decisions resulting in reported AGL performance, and not grammatical rules. In the following section I describe the procedure and details of a representative subset of these experiments, as well as my modeling of using PARSER. The complete set of experiments can be found, and replicated using the supplementary Juptyer notebook "PARSER experiments".

4.3.1. EEG potentials associated with Artificial grammar learning in the primate brain

Reference study: Attaheri, A. *et al.* EEG potentials associated with artificial grammar learning in the primate brain. *Brain Lang.* **148,** 74–80 (2015).[23]

Summary: In this paper, the researchers did experiments on two Rhesus macaque monkeys. The experiments were conducted using an auditory artificial grammar learning task. The monkeys were first trained using an artificial-grammar as depicted in Figure 2. After the training phase they were tested on familiar strings, novel but grammatical strings and novel violating strings. The study shows that certain ERP (event-related potential) components are modulated to a greater extent when there was a stimulus that contained a violation than when it did not. It is claimed that both the results reported here and the results reported from a behavioral study using eye-tracking cannot be explained by simple learning strategies.

Stimulus:

Exposure	**Consistent**	**Violating**
acfc, acfcg, acgf, acgfcg, adcf, adcfc, adcfcg, adcgfc	acfcg, acgfc, adcgf, adcfc, acgfc, acfcg, adcgf, adcfc	adfcg, adgfc, acfcf, adcgc, agfcd, afcgc, agcfc, acdfc

Table 1. The string-set used in experimentation. Collected from the paper by Attaheri, A. et al.

Goals: Initially there was a single *quantitative* goal in replicating the experiment in this study. The goal is to show that the strings presented during the habituation phase, can be learned by a simple learning strategy, such as PARSER. I hypothesize that the average PARSER model trained with similar stimuli as the monkeys in the original experiment will be able to learn to distinguish the

consistent and violating strings at a level higher than chance. To test the model's performance in achieving this goal, a more machine-learning performance approach is taken. Given that my hypothesis is correct, and that the average PARSER model indeed can learn to discriminate between the two classes: I wish to discuss the relation between the observed ERPs in the study and the inner workings of the model. If there is a strong relationship between the two, then the model could act as a violation predictor for future experiments.

Experiment translation: The experimental procedure in the paper is described as follows: The Macaques were exposed to familiarization artificial grammar sequences for 3 minutes. The monkeys had been pre-trained to conduct a visual fixation task during sound simulation. The training phase of the experiment consisted of exposing the animals to familiarization strings that were consistent with an artificial grammar, for 30 minutes. The presentation of a single item required the monkey to first fixate on a point for 500ms, then the string is audially presented (2665ms), and thereafter, a longer than 4500ms inter-trial-interval before the fixation point re-appeared and the process began anew. Following a completed training phase, a testing phase was imitated and lasted for 30 minutes. During testing, EEG data was collected and paired with the presented string. The EEG data recorded at the presentation of a consistent string was then compared to the EEG data recorded at the presentation of a violating string.

For this experiment, the procedure described above need translating into python-code so that the model could be trained in a similar scenario. The presentation of a string takes on average (500ms fixation + 2665ms string presentation + 4500ms inter-trial interval) 7.6 seconds and the training session lasts for 30 minutes. It is therefore calculated that the entire training phase consisted of the presentation of roughly 225 strings (7.6s ≈ 8s. (60/)8 *30). 225 strings were then selected at random (repetitions allowed) and then fed to the model as a concatenated string. This procedure was repeated for a set of 1000 models so that I could calculate how the average model responds to the stimuli. A grammatical inference method was used to test whether or not the models had learned the grammar. In the EEG data, the authors were specifically searching for ERPs that correlate with violating strings; they consider that the classification of a string being violating or consistent being a consequence of a missing or a lowered ERP response. Therefore, in analyzing the performance of the models in this experiment, I consider a *True Positive (*TP) to be a string that has been correctly classified as a violating string. A *False Positive* (FP) is a consistent string classified as violating, a *False Negative (*FN) is a violating string classified as consistent and a *True Negative* (TN) a consistent string classified as consistent. The reporting of these performance measures would answer the first goal of this experiment.

In the paper, they hypothesize that there will be a modulation of several ERP components related to the violating sequence in a string. This becomes the equivalent of hypothesizing that the model will

report back immediately after a sequence believed to be violating has occurred. Therefore, to investigate the second goal of this experiment, I train the model and look deeper into what caused it to classify the strings as violating or consistent. While this certainly would be interesting to do with multiple models for comparing the violating sequences in much the same manner as they do in the paper, it has only been done using a single model which is known to be able to make the classification well.

Results: First the 1000 models were exposed to the exposure string sets before they were all tested using a grammatical inference method. Evaluating the models using common performance measures from the field of machine learning shows that the average model is indeed is sensitive to the detection of violating strings. The mean accuracy of the trained models lies at 84% meaning that the average model has a high number of correct predictions (both of violating and consistent). In this experiment, it could be argued that due to the low number of data points, and a reasonably well-distributed data set, accuracy would be an accurate enough measure on the performance of the data. However, to account for a potential accuracy paradox, the performance measures *precision*, and *recall* is also reported in Table 2.

	Violating	Consistent	0.8425	Accuracy
Violating	7.2	1.09	0.892	Precision
Consistent	0.8	2.91		
Population 12	0.9			
	Recall			

Table 2. Confusion matrix showing the average performance measures of a 1000 models.

Discussion: The model is not a perfect learner of this string-set, while some models learn most of the strings, others fail to learn much. The average accuracy of the trained models shows however that the average model is capable of distinguishing violating strings from consistent strings on a level far above chance level. However, what does the model learn, and is it in relation to the grammar and the grand average ERPs that the monkey experiment report on? To investigate this a more detailed view of the results of the testing procedure was conducted. It would have been interesting to look at this type of behavior for multiple models, however, due to time constraints, only a single model which is known to be able to make correct classifications is considered. The strings that were classified as violating are reported in Table 3. In this table the column *ID* is referring to the IDs of the strings in Figure 1, column *word* is the full word being tested, *completed* is what parts of the string the model considers consistent, and the column *failed* is what parts of the string the model considers inconsistent or violating. As one can see from comparing the strings in Table 3 with the strings in Figure 1, the model is capable of picking up on all but one of the violations reported in red in Figure 1. This string, C3-1,

should according to the red markings in Figure 1 also have the violation *[g f]*. However, if one considers the language in which these strings are taken from before they are modified (see Figure 2), then one can see that this transition is in fact not an illegal move if one considers single transitions. Since the model is set up to learn these single transitions, it does not consider the transition *[g f]* as a sign of a violation. This is also the reason for why the string C2-1 is not in Table 3. The model considers the C2-1 string as consistent because it considers the transition *[c f]* legal. It would have been very interesting to have the average ERP responses to each of the individual strings, instead of a report on the grand average ERP which is made from all comparisons between violating and consistent strings. It would have been exciting to compare the C2-1 response in the monkey to the model's response. If the recordings during the C2-1 string did not show a difference in ERP, it could have acted as a reliable indicator that chunking or segmentation is the mechanism which accounts for grammar learning.

ID	Word	Completed	Failed
C1-1	adfcg	a [ ] fcg	[d f]
C1-2	adgfc	a [ ] gfc	[d g]
C2-2	adcgc	adc [ ]	[g c]
C3-1	agfcd	[ ] gf [ ]	[a g][c d]
C3-2	afcgc	[ ] fc [ ]	[a f] [g c]
C4-1	agcfc	[ ] [ ] cfc	[a g] [g c]
C4-2	acdfc	a [ ] [ ] fc	[c d] [d f]

Table 3. A model's inner workings when being tested. The column 'Word' is the full word being tested. Column 'Completed' is what the model managed to consider grammatical. Column 'Failed' is the parts of the word the model marked as violating

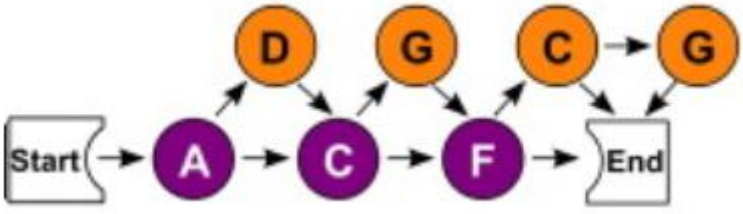

Figure 2. The artificial grammar utilized in the paper by Wilson et al.[23]

Violation					
C1	A	D	F	C	G
C1	A	D	G	F	C
C2	A	C	F	C	F
C2	A	D	C	G	C

Violation					
C3	A	G	F	C	D
C3	A	F	C	G	C
C4	A	G	C	F	C
C4	A	C	D	F	C

Figure 1. The set of violating strings used for testing the macaque monkeys. Extracted from the supplementary data from the article.

4.3.2. Rule learning by zebra finches in an artificial grammar learning task: which rule?

Reference study: van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B. & ten Cate, C. Rule learning by zebra finches in an artificial grammar learning task: which rule? *Anim. Cogn.* **16,** 165–175 (2013)[78]

Stimulus:

Experiment 1			
Go	**No-go**	**Probeset 1**	**Probeset 2**
ABA, BAB	BBA, AAB, BAA, ABB	ACA, CAC, ABBA, BAAB	ABAB, BABA, AABB, BBAA
Experiment 2			
Go	**No-go**		
ABA, BAB, AABA, BAAB	BBA, AAB, BAA, ABB, ABAB, BABA, AABB, BBAA		

Table 4. The stimulus created from an artificial grammar. From van Heijningen et al.'s paper.

Summary: In this paper, researchers did probing experiments on zebra finches. They used an AGL for creating stimulus and tested the zebra finches in a go/no-go experiment. The birds were trained using a go/no-go operant conditioning procedure, where they were rewarded with food. The training was conducted in the following way: When the birds pecked a switch in the left side of the cage, a sound was played, and a switch at the right side of the cage was activated. In fifty percent of the cases when the bird pecked the left sensory, a go stimulus was played, in the other fifty percent, a no-go stimulus sound was played. The bird was rewarded with food if it proceeded to peck the right sensor after a go-stimulus had been played, and was 'penalized' by having the light switched if it pecked the right sensor after the no-go stimulus was played. The birds were first trained in the go/no-go procedure using natural zebra finch song samples, where the birds needed to peck the sensor at least 75 percent of the times the go stimulus was played, and peck the sensor at max 25 percent of the times the no-go stimulus was played. After the birds had reached this criterion two days in a row, the, and the natural song stimulus was switched out with stimulus from the artificial grammar. After the birds had reached the same performance criterion on the AG stimulus, the birds were tested/probed using two probe-string sets with transitions based on specific rules. The study suggests that discrimination was dependent on the presence or absence of repeated A- and B- elements and that one bird was able to generalize the learned rules to a new symbol. In a second experiment, continued training the birds but with additional go and no-go strings based on a new rule. In this experiment, only two birds reached the performance criteria, and the authors claim that the birds differed in their discrimination strategies. The authors conclude that they have demonstrated bird's ability for rule-learning.

Goals: The goal of replicating this study is to investigate whether the performance of the zebra finches could have been explained using a chunking mechanism. By attempting to replicate the experimental procedure described in the paper, I hypothesize that PARSER models are capable of learning the string set resulting in a similar performance as the zebra finches.

Experiment 1 translation: Much of the experimental procedure has been detailed in the summary of this article. Here I describe how it was implemented in python. An important note, there is a clear difference in how many trials a primate from the experiment described in chapter 4.3.1 and a zebra finch in this study, use to reach the training criterion. This suggests that there is a clear difference in the memory capacities of primates and birds. This is something I wanted to account for in the PARSER models representing the zebra finches. Therefore, I assume that the number of percepts a bird can perceive together is less than that of a human, and change the max number of percepts perceived from three to two. I also assume that a zebra finch forgets faster and is more prone to perceptive interference than primates, and therefore change the forgetting weight to 0.1 and the interference weight to 0.01.

In the paper, eight out of ten birds managed to learn the discrimination. I, therefore, trained eight models on the go string set until they reached the performance criterion of maximum twenty-five percent no-go stimulus classified as go, and at a minimum of seventy-five percent go stimulus classified as go stimulus. A problem here is that a PARSER model can reach this performance criterion very fast (within two cycles). To avoid this, I also added the criterion that the model needs to have gone through at least three hundred perception cycles. The process of meeting the criteria was repeated twice to simulate the experiments criteria that the birds had to reach a confident performance two days in a row. After training was completed, all of the models were tested on all of the strings in probe-set 1. Simulating probing was done by first letting the model be trained on the probing string, then testing the model on that string. This is as close to the original experiment one can get. Each of the probing strings was presented to the models at least 40 times. After probing the models with probe-set one, the same procedure for probing was done for testing the models on probe-set 2. The models' performance was recorded after being probed with probe-set one and recorded again after being probed with probe-set two.

Results: All of the trained models shows a perfect classification of the go stimulus. Some of the models occasionally classified the no-go stimulus wrongly as go stimulus. For experiment one, the paper only reports the average response probabilities to the stimulus for the eight birds that did learn to correctly discriminate between go and no-go, therefore, to compare the models' performance to the bird's performance I completed the same averaging of my results. The authors in the paper also included G-tests of independence between training stimuli and between training and probe stimulus. My analysis did not include such an independence test. Figure 3 shows the average responses reported

in the paper. Figure 4 shows the average responses reported in my experiments with the PARSER model.

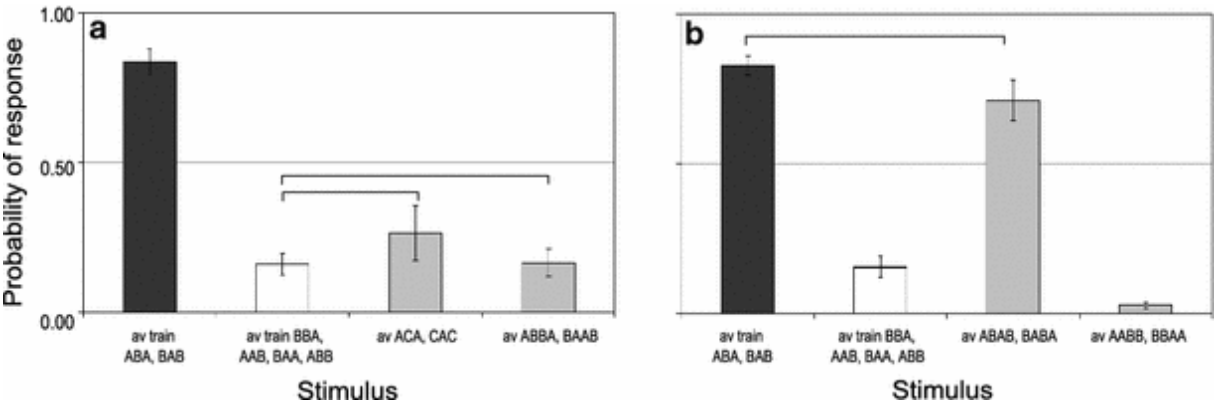

Figure 3. Averaged response probabilities to the stimuli (n = 8) during the first (a) and second (b) probe test (±SEM) of Experiment 1. Black bars indicate positive training stimuli; white bars negative training stimuli and gray bars probe stimuli. Taken from the results from van Heijningen et al.'s paper.

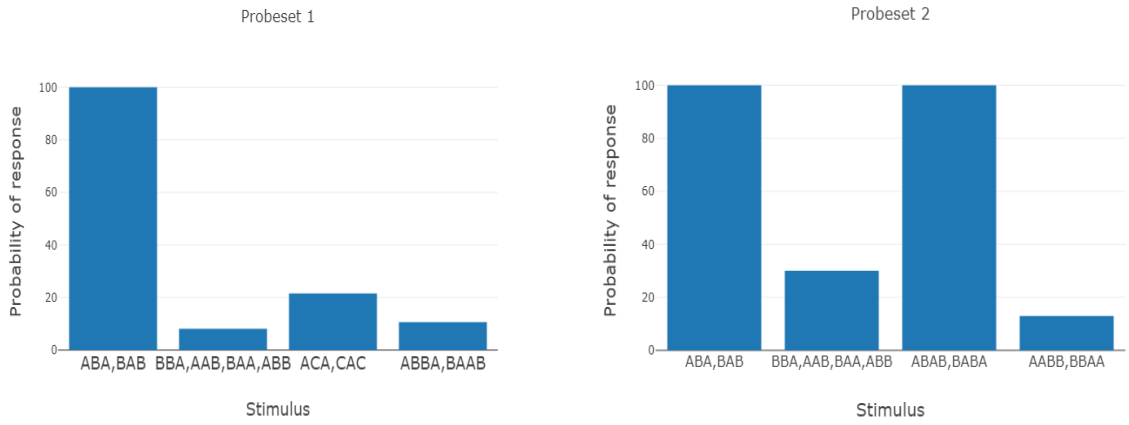

Figure 4. Average responses from the PARSER model after being probed with probe-set 1, and then being trained on probe-set 2.

Unfortunately, the authors do not provide the data from individual birds in this experiment, for this reason, a replication of the same graph of average responses was the closest I could get to compare the results. Figure 3and Figure 4 are remarkably similar. The difference between them is that all of the PARSER models learned to correctly classify the go stimulus. The (ABAB, BABA) category in probe-set two, has also been learned by all of the models. This, however, is not surprising due to how the chunking mechanisms and the grammatical inference method works. ABAB and BABA have the same set of transitions as ABA and BAB, so if the model has learned ABA and BAB, then naturally ABAB and BABA will be marked as grammatical as well.

Discussion: The results show that a chunking mechanism is capable of showing similar performance as the zebra finches' performance reported in van Heijningen et al.'s paper[78]. In the paper, they claim that the stimulus set used for training show that zebra finches can distinguish strings where

single element positions or counting of A or B elements could be used as cues for discrimination. The model shows similar performance. However, the authors go on to elaborate on the results of the probe tests show that the birds are using specific rules to discriminate between the stimuli. While it can be agreed that one can describe the performance behavior of the birds by describing what rules they are sensitive to, I argue that such an explanation is shallow if the goal is to explain the cognitive mechanisms of the birds.

4.3.3. Other studies and experiments.

A number of other experiments were conducted using PARSER. These experiments have not been replicated with as much detail as the two experiments reported previously. These experiments have been conducted to see if the PARSER model can learn to distinguish the grammatical and ungrammatical strings from a number of string-sets. It is important to stress that the

Wilson et al. (2015): Auditory sequence processing reveals evolutionarily conserved regions of frontal cortex in macaques and humans[79].

Quick summary: In this study, the authors use fMRI and an AGL paradigm to compare the brain regions of humans and monkeys thought to be responsible for initial syntax processing. The study shows evidence that regions known to be associated with syntax processing, have functional counterparts in monkey brains. The monkeys are tested on their ability to detect violating strings.

Stimuli:

Exposure	**CorrectTest**	**ViolatingTest**
acf, acfc, acgf, acgfc, adcf, adcfc, adcfcg, adcgf, acgfcg	acgfc, adcfcg, acfcg, adcgfc	afgcd, afcdgc, fadgc, dcafgc,

Table 5. The string set from Wilson et al.'s paper[79].

Experimentation: I wished to know whether the PARSER model could learn to distinguish the two categories ‘CorrectTest’ and ‘ViolatingTest’ from only being exposed to the ‘Exposure’ strings shown in the stimuli-table above. To do so, I trained a thousand PARSER models on 400 strings from the ‘Exposure’ string-set. The 400 strings were selected in a random order. PARSERs default parameters were used. As with the monkeys, the model was evaluated on its performance to detect violating strings.

Results: The average PARSER is very much able to distinguish the two categories. All of the models learn to correctly classify the violating strings. However, on average, one of the consistent strings get classified as a violating string as well.

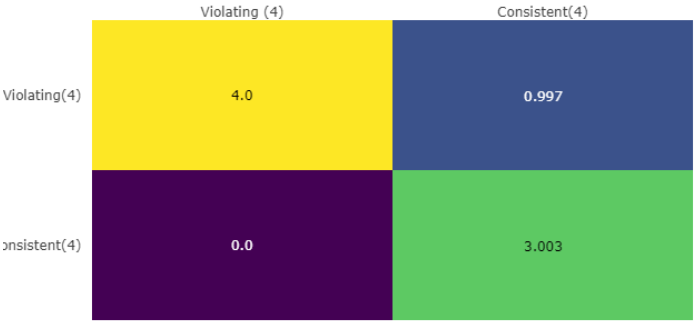

Figure 5. The average performance of 1000 PARSER models being exposed to stimuli from Wilson et al.'s paper [79].

Wilson B, Smith K, Petkov CI (2015): Mixed-complexity artificial grammar learning in humans and macaque monkeys: evaluating learning strategies[80].

Quick summary: In this study the authors test humans and Rhesus macaque monkeys are trained on a mixed-complexity auditory artificial grammar. The grammar contains both adjacent and non-adjacent relationship. The participants (both monkeys and humans) were then tested on consistent sequences and on sequences that contains specific violations. Both species showed a sensitization to the adjacent relationships as well as to the statistical properties of the sequences. However, there was no significant sensitivity observed to the non-adjacent relationships in the macaques, while a small subset of the humans were sensitive to this property. The result suggests that both species are comparable in sensitivity to adjacent AG relationships, but macaques are less sensitive to the non-adjacent AG stimulus.

Stimuli:

Exposure	**CorrectTest**	**ViolationTest**
acf, acfcg, acgf, acgfcg, adcf, adcfc, adcfcg, adcgfcg	acfcg, adcfc, acgfc, adcgf,	adcfg, adfcg, adgcf, adgfc, agcfc, agfgc, agdcf, agfdc

Table 6. The string set from Wilson et al.'s paper[80].

Experimentation: The string set used in this experiment is somewhat similar to the one used in the previous experiment described above. The procedure is the same: I trained 1000 PARSER models on the 'Exposure' string-set to test its ability to detect the violating strings in the set. 400 strings from the 'Exposure' string-set, in random order, were used as input to the models. PARSERS parameters were left untouched, meaning they were set at default.

Results: The average PARSER model shows a great ability to detect the violating strings. When it comes to the consistent strings, however, the average model fails to classify one of them correctly.

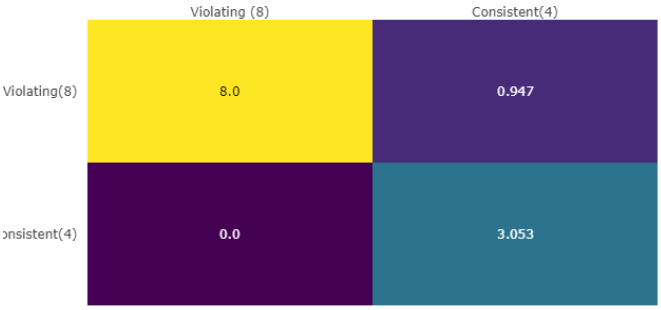

Figure 6. The average performance of 1000 PARSER models being exposed to stimuli from Wilson, Smith and Petkov.'s paper[80].

Wilson et al. (2013): Auditory Artificial Grammar Learning in Macaque and Marmoset Monkeys[81].

Quick summary: In this study a number of monkeys were habituated to an artificial grammar with a forward branching structure. The authors argue that such a forward branching structure in an artificial grammar is designed to model specific properties of the non-deterministic nature of symbol transitions in natural language and animal song. After the habituation phase, the marmoset monkeys were able to differentiate between grammatical test strings and violating strings. According to the authors, the macaque monkeys demonstrate a more complex artificial grammar learning, and that this is evidence for macaques handling deeper levels of AGL. However, no mention on what constitutes deeper levels of AGL is given. The acknowledge that their results fail to rule out a reliance on simple strategies, but conclude the paper suggesting that marmoset and macaques could potentially be used as model systems to study aspects of AGL at a neural level.

Stimuli:

Hab	Fam	Novel	ViolbA	ViolnbA
acf, acfc, acgf, acgfc, acgfcg, adcf, adcfc, adcfcg, adcgf	acgfc, adcfcg	acfcg, adcgfc	afgcd, afcdgc	fadgc, dcafgc

Table 7. The string set from Wilson et al.'s paper[81].

Experimentation: In this experiment, the 'Hab' string-set is the exposure string set, the sets 'Fam' and 'Novel' make up the set of grammatical strings, and 'ViolbA' and 'ViolnbA' make up the set of ungrammatical strings. I trained 1000 PARSER models on the 'Hab' string-set. I used a random sequence of 400 strings to do so. Then I tested the models on their ability to correctly classify ungrammatical strings as violating and grammatical strings as consistent. PARSERs parameters were set to default.

Results: The average PARSER model is to a high degree able to correctly classify violating and consistent strings. Since the number of average classification of violating strings is four and there are only four violating strings, it yields that all of the thousand models can correctly classify the violating strings. There is, however, some of the consistent strings that are being wrongly classified as violating.

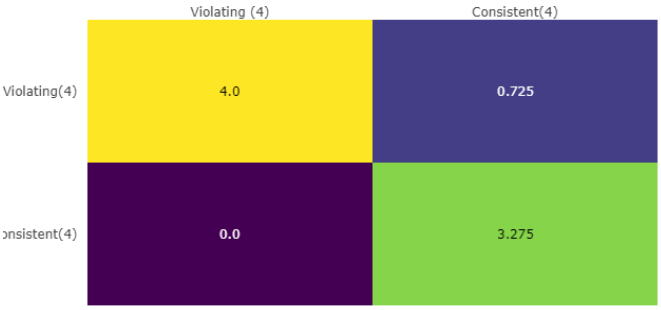

Figure 7. The average performance of 1000 PARSER models being exposed to stimuli from Wilson et al. (2013)'s paper[81].

Saffran et al. (2008): Grammatical pattern learning by human infants and cotton-top tamarin monkeys[24].

Quick summary: In this study, human infants and cotton-top tamarin monkeys are trained and tested on their ability to learn grammatical structures. The infants and the monkeys were subject to comparable experimental designs, so that their results could be compared. The results show that the human infants can quickly pick up on complex grammatical structures (non-predictive language), while the tamarins could only learn the simple grammatical structures (predictive language).

Stimuli:

PLS	NPLS	GTS	UTS
bcd, bcjd, bkcjd, bkcd, bcdcj, bkcdcj, bcdc, bkcdc	bcd, bcjd, bkcjd, bkcd, kcjd, bkjd, kcd, bjd	bcd, bkcjd, bkcd, bcjd	bckd, bjcd, bkd, bcjkd,

Table 8. The string set collected from Saffran et al.'s paper[24].

Experimentation: The exposure string-set used in this experiment is made from the 'PLS' and 'NPLS' string-sets as seen in the table above. The grammatical strings for testing are the 'GTS' string-set, and the ungrammatical strings for testing are the 'UTS' string-set. 1000 PARSER models were exposed to the predictive language and another 1000 PARSER models were exposed to the NPLS set. 400 strings were randomly selected from the exposure sets to serve as input for the models. The models were then tested to see if they could correctly classify the grammatical strings from the 'GTS'

string-set and the ungrammatical strings from the 'UTS' string set. PARSERs default parameters were used.

Results: Exposing the model to the PLS string set shows that the model is capable of learning to correctly classify both violating and consistent strings with a high accuracy. However, when exposed to a non-predictive language the correct classification of consistent strings severely drops, and the model classifies grammatical strings as violating more often than not. This suggests that the PARSER model might not be able to correctly learn structures lacking predictive dependencies. This in line with the original experiments results.

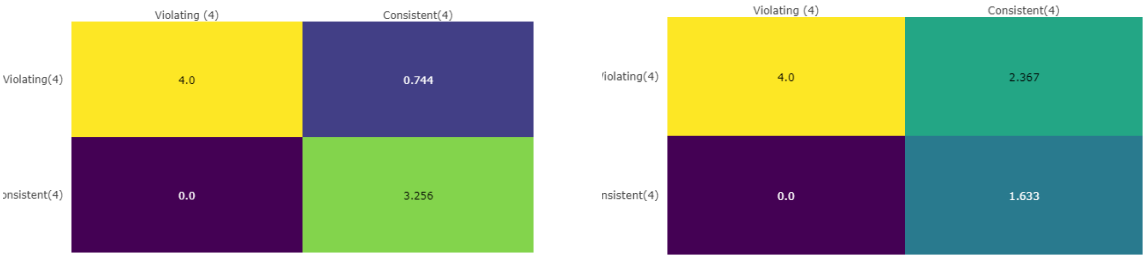

Figure 8. The average performance of 1000 PARSER models being exposed to stimuli from Saffran et al.'s paper24. The left matrix shows the results from the models being exposed to a predictive language (PLS exposure) and the right matrix shows the results from the models being exposed to a non-predictive language (NPLS exposure).

Endress et al. (2009): The apes' edge: positional learning in chimpanzees and humans[82].

Quick summary: In this study chimpanzees and human adults are exposed to six auditory sequences created from simple temporal rules. The rules for creating the training stimulus follow the following pattern: an A item always precede a B item, and the edges of the strings edges are populated by an X item. Their results suggest that human and chimpanzees both have a cognitive mechanism to encode positional information from sequences automatically. A string was classified as violating if the monkeys reacted to the sound of it.

Stimuli:

Habituation	**Test**
XABXXX, XAXBXX, XAXXBX, XXXABX	AXXXXB, BXXXXA, XXABXX, XXBAXX, XXAXBX, XXBXAX

Table 9. The string set collected from Endress et al.'s paper.

Experimentation: This experiment did not explicitly test grammatical and ungrammatical strings. The strings in the 'Test'-string set are all unique as they either have violations at specific transitions. I trained 30 PARSER models on 400 randomly selected strings from the 'Habituation' string-set. Then I

tested all of the models on all of the strings in the test set. PARSERs default parameters were used. The models were tested on their ability to 'reject' a string, meaning that they would classify the string as violating.

Results: The results are shown in the Figure 9. All of the table, except for the last column (PARSER) is taken from Endress et al.'s[82] paper. It highlights if the strings have regularity chaining and if the string respects the positions of the X symbol in the training set. The respected positional means that strings start and end with an X, and regularity chaining is whether the A occurs before B. It is important to note that since the model is set to detecting violating strings, the numbers reported in Figure 9 are the number of models that have marked the strings has violating. The flipped (meaning the number of models minus the number of models that classified the string as violating) is the number of models that classified the string as grammatical. This shows that most of the models do not detect the violations in the string, and therefore classify the strings as grammatical. The violation detection of the model seems to fall in between the performances' of the humans in Exp.2a and the chimpanzees in Exp.1, except for the string XXBAXX which is classified as violating by all 30 trained models. The reason for this is rather simple, none of the strings in the 'Habituation' string-set contains the sequence BA, which results in the model not having the data to learn it.

Number	Item	Regularity Chaining	Respected positional	Exp.1	Exp.2a	Exp.2b	Exp.3	PARSER
1	AXXXXB	Yes	No	8/15	23/30	20/30	20/30	9/30
2	BXXXXA	No	No	6/15	25/30	21/30	23/30	1/30
3	XXABXX	Yes	Yes	2/15	4/30	4/30	9/30	7/30
4	XXBAXX	No	Yes	5/15	17/30	4/30	8/30	30/30
5	XXAXBX	Yes	Yes	1/15	8/30	4/30	7/30	10/30
6	XXBXAX	No	Yes	2/15	17/30	10/30	15/30	10/30

Figure 9. A table from Endress et al.'s paper[82]*, with an added column showing PARSERS results. The 'Item' column is the string being tested. 'Regularity chaining' is an indication of whether the string follows the pattern: A comes before B. 'Respected positional' is whether the string follows the pattern: Starts and ends with X. 'Exp.1' is the number of times chimpanzees reacted to the strings. 'Exp.2a' is the number of times a human rejected the string in an experiment using noises as symbols. 'Exp.2b' is the number of times a human rejected the string in an experiment using human speech syllables as symbols. 'Exp.3' is the number of times a human rejected the string in an experiment using the same symbols as in 'Exp.2b', using more exposure strings. 'PARSER' is the number of times a different PARSER models classified the string as grammatical.*

5 Results

The goal of the experiments presented in chapter 4.3 has been to investigate whether a computational model describing the mechanisms suggested to account for word segmentation in humans, would be able to account for animals performance in AGL experiments. To do this, two experiments from the literature were replicated in detail in silico. And the model was trained and tested on a number of other string-sets used in the literature to see if it could be able to distinguish between grammatical and ungrammatical stimulus. Since the results from the two in-depth experiments have already been presented in sections 4.3.1 and 4.3.2, I only quickly summarize them here. In the first experiment, a replication of an EEG study on monkeys[23] shows two things. Firstly, the results show that the average PARSER model trained on the same stimulus as presented in the paper is indeed capable of distinguishing between the grammar produced stimuli, and stimulus that has violating transitions. With an accuracy of about 84%, the model can do this at a level far above chance. Secondly, the performance of a model known to be able to make correct classifications shows that there is a correlation between what the model predicts as being a violating transition and where the ERPs from the EEG data occur. The second experiment, replicating a study on rule-learning in zebra finches[78], shows that the PARSER model is able to generalize its learning to novel stimulus with a performance similar to that of the zebra finches. The authors argue that some of the birds able to transfer their rule-knowledge to a new set of symbols. Because of PARSERs internal mechanism, such a transfer of rules should not be possible. However, the results show that the PARSER model performs similarly to the birds in this task. The reason for this is that the model picks up on the chunks in the stimulus, and when the stimulus' symbols change and are exposed enough times to the model, there exists the probability of the model learning it. Meaning that transfer knowledge is not necessarily needed for performing well in transfer tasks.

In chapter 4.3.3, several other string-sets were used to test the performance of the PARSER model. Due to time constraints, I presented the model with the string sets in a procedure as to see if the model could learn to distinguish between grammatical and ungrammatical strings, and not to replicate the experiments from where the strings-sets were collected. The string-set provided by Endress et al[82] was not fit for testing grammatical and ungrammatical strings and was for this reason experimented with in a different way. All the other string sets, Wilson et al[79], Wilson et al[80], Wilson et al[81], and Saffran et al[24] were tested in the same fashion. They were trained on an exposure string-set, then tested on both grammatical and ungrammatical strings. Their performance was recorded in common machine learning performance measures, reported in Table 10. In this table, the true positives are the number of ungrammatical strings classified as ungrammatical, false positives are the number of grammatical strings classified as ungrammatical, true negative is the number of grammatical strings classified as grammatical, and false negative is the number of ungrammatical strings being classified as grammatical. Accuracy is the proportion of correctly classified strings in the data. Precision is an

expression of the number of strings PARSER says are correct, that actually is correct, and recall is an expression of the model's ability to find all of the relevant (in this case grammatical) strings in the data. The F1 score is a harmonic mean of recall and precision. The goal, generally speaking, is to maximize the F1 score. The string-sets that came from Wilson et al[79],Wilson, Smith, Petkov[80] and Wilson et al[81] were all generated from the same finite state grammar but varied in what strings were presented in the habituation phase, and what strings were used for testing. However, due to them all being generated from the same grammar, PARSER learned to distinguish the grammatical and ungrammatical strings with a performance similar across all of the experiments. Training the models on the 'PLS' string-set from Saffran et al.[24] (Table 8) paper resulted in the model being able to correctly learn to classify both grammatical and ungrammatical strings. However, when the model was trained on the 'NPLS' strings from the same set classification of grammatical strings severely dropped. This suggests that the PARSER might not be able to learn to classify grammars lacking predictive dependencies.

Measures (Detect violating string)	**Wilson et al**[79]	**Wilson et al**[80]	**Wilson et al**[81]	**Saffran et al (PLS)** [24]	**Saffran et al (NPLS)** [24]
True positive	4	8	4	4	4
False positive	0.997	0.947	0.725	0.744	2.367
True negative	3.003	3.053	3.275	3.256	1.633
False negative	0	0	0	0	0
Accuracy	0.8754	0.9211	0.9094	0.9070	0.7041
Precision	0.8005	0.8942	0.8466	0.8432	0.6282
Sensitivity/recall	1	1	1	1	1
F1	0.8892	0.9441	0.9169	0.9149	0.7717

Table 10. A number of performance measures for the experiments conducted in chapter 4. The numbers reported here are the same performance measures as seen in chapter 4. These are on the basis of PARSER detecting violations.

Something that I would have to remark upon here is that when the PARSER model is set to detect violating strings the number of true positives is the number of violating strings that are classified as violating. This means that when PARSER is making a judgment on the grammatical well-formedness of a string, it is classified as violating if it does not manage to build the test string from its vocabulary. This implies that there is an inherent bias for classifying strings as ungrammatical or violating. To account for this, I also provide the same measures for a flipped version of the performance. Meaning that a true positive in Table 11 is the number of grammatical strings classified grammatical strings, false positive is the number of violating strings classified as grammatical, true negative is the number of ungrammatical strings classified as ungrammatical, and false negative is the number of ungrammatical strings classified as grammatical. As the general goal is to maximize the F1 score, the performance shown when detecting violating strings (Table 10) is better than detecting grammatical strings (Table 11). However, both versions show the ability to correctly classify both categories. But because of the bias for detecting violating strings, it could be argued that the performance shown from detecting grammatical strings is a more accurate description of the actual performance of the model.

Measures (Detect grammatical strings)	**Wilson et al**[79]	**Wilson et al**[80]	**Wilson et al**[81]	**Saffran et al(PLS)**[24]	**Saffran et al(NPLS)**[24]
True positive	3.003	3.053	3.275	3.256	1.633
False positive	0	0	0	0	0
True negative	4	8	4	4	4
False negative	0.997	0.947	0.725	0.744	2.367
Accuracy	0.8754	0.9211	0.9094	0.9070	0.7041
Precision	1	1	1	1	1
Sensitivity/recall	0.7508	0.7633	0.8188	0.8140	0.4083
F1	0.8576	0.8261	0.9003	0.8975	0.5798

Table 11. Flipped performance measures for the experiments conducted in chapter 4. The numbers reported here is if the model is detecting grammatical strings.

A final experiment using a string-set from Endress et al.'s[82] paper was conducted. In this string-set, there was a clear separation between grammatical and ungrammatical strings. However, the strings were built from two simpler rules than most other grammars are made from. The strings were generated from the rules; A must occur before B and strings must start and end the symbol X. I trained thirty PARSER models on an exposure-string set and tested them on all of the test-strings. In the test-set, two strings were consistent with the rules, and four had violations on one or both of the rules. The performance of the models is reported in Figure 9. In this experiment, the models were tested on their ability to detect violating strings. Most of the models were classifying the majority of the strings as grammatical, not violating, after being trained on a sequence of 400 exposure strings.

6 Discussion

This chapter aim is to discuss the various topics, questions, and goals presented in this thesis. To do so, I first discuss the strengths and weaknesses of this research, with a focus on the PARSER model and the experiments conducted. Then a more general discussion on the findings from the experimentation reported in chapter 4.3 and chapter 5 is done through a comparison with the literature discussed in chapter 2 and 3, as well as goals presented in the introduction. And finally, to complete this chapter, I broaden the scope of the discussion to the implications of this research.

6.1. Strengths and weaknesses of experimentation and modeling

One of the goals of this thesis has been to show that there are cognitive mechanisms other than rule-learning that are capable of producing similar performance reported in animal AGL experiments. In the review of the AGL literature, it was discussed *what* participants are possibly leaning in these experiments and *how* participants are learning *what* that is being learned. In chapter three a number of computational models describing cognitive mechanisms representing the *how* of AGL were detailed. One of these models, PARSER, was trained in the same procedure as a number of animal AGL experiments. The results from these simulated experiments clearly show that a chunking model is capable of replicating the performance of animals in AGL experiments. However, as it is with any model or experiment, they do not come without a set of limitations.

The first topic to address is the appropriateness of modeling artificial grammar learning using a chunking model, or more specifically a word-segmentation model. As mentioned in chapter 4, PARSER, and any word-segmentation model, is without a method for grammatical inference unfit for modeling AGL. In both a word segmentation task and an AGL task, the participant is presented with an input stream consistent with a grammar. The difference is only what the participant is asked to do after exposure. In an AGL task, the participant is asked to classify the well-formedness of novel strings both consistent and in violation with the grammar used to produce the exposure data. In a word segmentation task, the participant is asked to classify what constitutes words in the exposure data. The difference lies in the word-segmentation model not being asked to generalize its vocabulary to a grammar, but only to induce what makes up words in the exposure set. Therefore, since the learning mechanisms underlying both an AGL task and a word-segmentations task is arguably the same, and it is the learning mechanism that we wish to model, it becomes appropriate to use a word-segmentation model to model AGL performance.

The second topic that I would like to address is the added grammatical judgment method to the PARSER model. Perruchet and Vinter's[59] model did not directly address the concept of grammatical inference, but rather the concept of word-segmentation. Their criterion for assessing the model's performance is based on the trained model's vocabulary. A loose criterion is met if the words of the language have the highest strength in the vocabulary, and a strict criterion is met if the

vocabulary does not contain chunks nor part-words, but only contains the words of the language. Implicated from this is that the model, as it was presented by Perruchet and Vinter, does not yield any form of generalization to words it has never encountered. Using the loose and strict criteria leaves the model unfit for use in AGL tasks, as novels strings will not be part of the vocabulary. The addition of a grammatical judgment method, therefore, seems appropriate. However, even though I argue that the addition to the model is appropriate, how valid are the assumptions I infer in creating this method? And what is the cognitive plausibility of the three phase's *extraction*, *matching*, and *building*? The first assumption I made is that 'there should be no loss of information when that information is used to learn something new.' This assumption is made for the specific scenario of learning from data that does not contain any mistakes, meaning that all the data the model has access to, is correct information. Therefore, if one learns first a specific chunk, and then use that chunk to learn something new, then the new chunk has not lost the information contained in the first chunk. This is analogous to a first grader having learned first to write all the letters of the alphabet, and then go on to learn how to spell words. Then the information on how to write the letters is not lost as it is needed to write the words. Therefore, that assumption should be uncontroversial.

The second assumption stating that 'Memory is expensive, processing is cheap' is tricky to cognitively validate as I initially made the assumption from analyzing what exactly the PARSER model does when it computes its chunks. I took an optimization approach to analyze the model and argue from this that what the model attempts to do, is to save as much information, in as little memory as possible. The optimal solution for the PARSER model learning a language of ten words, in terms of memory, would be a vocabulary containing exactly the ten words of the language. However, the process for deriving these ten words is computationally costly because of how the model needs to compute inference weights, forgetting and other transformations of the vocabulary. Such optimization behavior is valid for all chunking models, and I would, therefore, argue that the second assumption is reasonable if the chunking mechanism is considered to be a valid cognitive mechanism.

The third assumption, which builds upon both the first and the second goes like this: 'A chunk is more than the chunk itself. A processed chunk is the combination of all the transitions between the symbols in that chunk'. To bring up the first grader learning to write again, this is the equivalent of writing an entire word. However, when the word is being written, it is not written in one go, it is being processed so that each individual letter will be written in sequence.

These three assumptions together make up what is the three phases *extraction, matching,* and *building.* But what is the cognitive plausibility of these phases? To discuss this, I would first like to stress that my area of expertise is not in neuroscience, biology nor psychology and that my knowledge of these fields should be considered limited. Therefore, debating whether or not this process is a good representation of a cognitive process becomes rather difficult. However, what I do have knowledge of

is the various arguments on cognitive plausibility from experiments and models of AGL. These arguments serve as rough guidelines for discussing the cognitive plausibility of the grammaticality judgment method. The extraction phase happens as the model is about to make a grammaticality judgment on a chunk. It begins with splitting the chunk into all possible smaller chunks contained in it. This should be considered a type of feature extraction of the input. Feature extraction of input is something that occurs at multiple places in the cognitive system in humans. An example of feature extraction is how the brain processes vision. Chang and Tsao[83] recorded the electrical signals in the brains of two rhesus monkeys while they were show almost 2000 artificially created human faces. These faces were made from 50 different features, such as skin color, nose size, and space between the eyes. The recordings resulted in a map of neurons firing in relation to the specific features the face had. While this type of feature extraction is the result of the propagation of electrical signals through neural networks in the brains temporal lobe, PARSER does this in a much more explicit symbolic way. However, while PARSER's feature extraction might not be in perfect correlation with how it happens in the brain, I would argue that on a computational and algorithmic level they are similar. One limitation in regards to the extraction method is that for every string that is being tested, the feature extraction is done on every chunk in the vocabulary. This, however, was a decision I made during implementation of PARSER. I could have added a separate vocabulary containing all of the features of all the chunks in the vocabulary. But that would have been done to optimize the algorithms in terms of speed and would have resulted in a too large deviation from the original model than what seemed appropriate. The *matching* phase is also a result of having implemented a symbolic version of the model. A symbolic comparison requires explicit matching, meaning that the matching phase could only fit with a computational explanation of the process. However, it could be argued that a matching phase would be somewhat equivalent to the firing of neurons that match both the input chunks and the memorized chunks. However, I consider such an argument only as speculation and more as an intuitive way of thinking about the matching phase. The final phase, *building*, is the attempt to re-build the input chunk from the matched chunks in the previous phase. If all of the extracted chunk from the input string can be found in the matched set, then the chunk is to be considered grammatical, and ungrammatical otherwise. The only way in which I could argue for the cognitive plausibility of the building phase is by comparing all of the three phases to the processes of an autoencoder. And then further argue that even though it is a symbolic model, it undergoes the same process as an autoencoder, and since the cognitive plausibility of autoencoders has previously been suggested, I will to use this argument. An autoencoder is a type of artificial neural network used to build efficient data encodings. The goal of an autoencoder is to learn to encode or compress data from input into a smaller piece of code, for then to use the encoding to generate something similar to the input data. The grammaticality judgment method does something similar, but instead of compressing the input it is processed to extract its features. Figure 10 shows an example of the three phases represented in an autoencoder structure.

The top graph in the figure shows an example of a consistent string being processed, leading to an output string that matches the input string. The bottom graph shows an example of a violating string been processed. The extracted vocabulary in the bottom graph does not result in a D-F transition, and therefore a failed match occurs, and the building phase results in a string that does not match the input string. It is important to note here that the 'vocabulary extraction' layer would possibly contain a whole lot more nodes than what is displayed here. In the end, I will only on the background of the autoencoder representation of the grammaticality judgment method call the entire process for cognitive plausible.

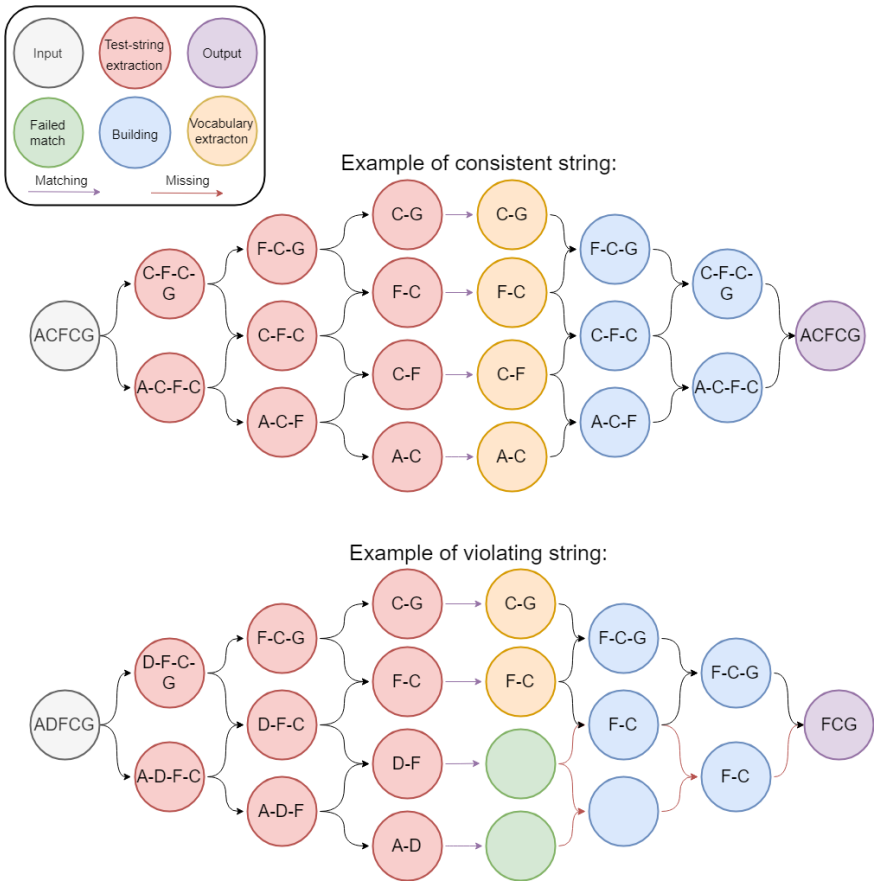

Figure 10. The three phases Extraction, matching and Building, represented as an autoencoder.

Another limitation of the grammaticality judgment method is that it does not utilize the chunk strengths that the vocabulary provides. Having an implementation that does use the strengths would give the potential to give a gradient on how grammatical the chunk is. Allowing for focus on more important parts of a chunk, instead of valuing all of them as equal. It could also be used to allow the matching procedure to take place between earlier extractions than the lowest level. Meaning that one could match whole chunks (ADFCG), and half chunks directly (F-C-G, A-D-F), instead of having to only match (C-G, F-C). This would greatly improve the grammaticality judgment method.

Another concern when it comes to the PARSER model is its parameter space. In the second experiment I conducted (chapter 4.3.2) I changed the parameters of the model in an attempt to account

for the difference in cognitive ability between monkeys, humans and zebra finches. The assumption that zebra finches cognitive abilities are less than that of monkeys and humans most likely hold true, however, it could be argued that such a change in the parameters might not be appropriate due to the unknown fitness landscape of the parameter space. However, I would argue that it would be less justified to run the experiments using the same parameters for memory capacities for these two species.

The zebra finch experiment replication also highlights another interesting point. The zebra finch experiment supposedly shows that (some) zebra finches are able to transfer their grammatical knowledge on to a new set of symbols. In the original experiment, the birds are trained using a go/no-go operant conditioning task, giving rewards upon correct behavior and 'punishment' for wrong behavior. During the test phase, there is no reinforcement of correct behavior and no punishment for wrong behavior. From this, it can be assumed that their argumentation that the birds are learning the grammar falls on the assumption that there are no strings from the probe test-set that are being learned. The birds are supposedly only learning from the strings which they were trained on, and then, because of the lack of reinforcement, stop learning during the test phase. However, PARSER does not possess any ability to abstract its vocabulary so as to generalize to a new set of symbols nor does it have any method for reinforcing the behavior. But still, it manages to perform equal to the eight birds that were trained and tested in the experiment. I argue that this shows that the birds are continuing to learn during the testing phase, and the results are mistakenly interpreted as transfer knowledge. I believe this highlights an issue with probing experiments, as they assume that the animals in question stop learning to use their knowledge when being tested, while it is all a continuous process for them. I am however unsure as to how one can go about changing the experimental design to accommodate this problem. But further investigation into the similarities and differences between PARSER and the birds would strongly benefit from having access to data from individual birds.

The 'other' experiments (chapter 4.3.3) shows that the PARSER model is indeed able to learn to distinguish between grammatical and ungrammatical strings after being exposed to grammar produced stimulus. To further explore the potential of the PARSER model as an explanatory model for AGL performance, future research would have to investigate exactly what types of stimulus the model is sensitive to, and to what degree this is comparable to the performance reported in animal studies. The mapping of what properties of stimulus the model is sensitive to is, however, not the purpose of this thesis. As the goal is to take a well-established model able to capture human AGL performance, and apply it on animal data to see if the model was able to capture animal artificial grammar learning as well.

6.2. General discussion on animal AGL research

One of the two main goals of this thesis was to investigate whether or not the animal AGL paradigm makes an unjust assumption that animals learn and encode abstract rules when presented with grammar produced stimulus. To reach this goal an examination of the animal AGL literature was conducted. The literature was examined with regards to *how* experiments are conducted, *why* these studies are conducted and *what* the studies conclude. According to this investigation, the problem lies not in *how* the experiments are conducted, but in the conclusions they draw from the experiments. Results indicating that animals are sensitive to the rules of an artificial grammar are not strong evidence per se, that animals are grammatically competent. And this indictment lies in what seems to be a deeply rooted assumption of rule-learning. However, being sensitive to rule-produced stimuli does not entail rule-learning. For rule-learning to occur, there needs to be a cognitive mechanism that is not just sensitive to the rules but also learns to encode them. None of the studies investigated in this thesis describe such cognitive mechanisms. The literature shows an apparent lack of reflection on what is to be regarded as behavior and what is regarded as mechanisms driving behavior. Being sensitive to an artificial grammar is behavior, but the sensitization is caused by a mechanism. With results only showing animals' behavior, one can only speculate as to what the mechanism is. But the animal AGL field does not speculate, it assumes. To make progress, it is necessary to arrive at specific hypotheses about what is causing behavior. And one needs to test such hypotheses by controlling for other mechanisms that could result in similar behavior. So far, there has been little emphasis on such an approach, leading to controversy as to how to interpret animal response behavior in animal AGL research.

In chapter 3.2 I introduced Marr's levels of analysis as a way to investigate complex processes. Marr's levels of analysis provide great value when it comes to thinking about cognition and the processes and mechanisms that it encompasses. The three levels, *computational, algorithmically and representational,* and *implementation*, each provide a piece to solve the overall puzzle of cognition. Evaluating contemporary behavioral animal AGL research using Marr's levels of analysis, reveals that its investigation is restricted the computational level, with little consideration to the two other levels. Having a single level explanations is, by all means, possible, as long as one does suggest that this is the whole picture. Marr, however, argued that for an explanation of a cognitive process to be sufficient, one needs to address the function of the computational level by answering why the function exists as well as what it does. Only answering one of these questions results in shallow explanations for a phenomenon. The problem is however not limited to the computational level. AGL studies using high-technology techniques such as EEG or fMRI falls, mostly, outside the category of behavioral studies. These studies are however also limited in the same sense as the behavioral experiments in the way that they only consider one level, the implementation level. Isolated, the implementation level

does not paint any more of the picture than the computational level does. Yet most papers present their research as if they could see the whole picture.

It becomes difficult to assess what exactly the reason for this is. Perhaps it is that the animal AGL field is still maturing and this line of thinking has yet to be implemented. Such a reason would, however, be somewhat unjust, as the AGL paradigm has its roots in human research and was created with the intention of deducing *what* the driving mechanisms were. Another reason that comes to mind is an insufficient distinction between behavior and mechanisms is. To illustrate this, I would like to revisit the example given in chapter 3.3 illustrating the difficulty of explaining the mechanisms of a sensor sitting on a window sill. To recap: You are tasked with figuring out what the sensor is doing. As you observe the sensor, you notice a small light switch off at the absence of sunlight. From this, you hypothesize that the sensor is reacting to the light. By continuous observation of the sensor over an extended period of time, you keep seeing the same behavior from the sensor: the presence of sunlight, the light switches on, the absence of sunlight, the light switches off. In the end, you believe that the amount of data that you have collected is enough proof that the sensor is a light detector. However, this proves not to be the case. The sensor is a temperature detector. And when the sun hits it or moves away from it, there is a small temperature change that causes the light to turn on or off. The sensor is absolutely sensitive to the sunlight, meaning that its behavior is reactive to the sunlight. But sunlight has nothing to do with the driving mechanisms for the state of the sensor's light.

Currently, the animal AGL paradigm finds itself in a similar position as the person observing the sensor. At some point, someone observed that a number of animals were sensitive to the stimuli from an artificial grammar, and then hypothesized that the animals are sensitive to the rules of the grammar. Through continuous testing of the hypothesis, it is induced that the driving mechanisms of AGL are rule-learning. There are two limitations to this approach. First, as I have explained above in the sensor example, the nature of the underlying mechanisms driving behavior need not have anything to do with how the stimulus that is presented is generated. An animal could show behavior indicating that it is sensitive to grammar produced stimuli, without the mechanisms even having to process the rules. Secondly, and on a more philosophical level, this type of research falls in the trap of inductive research, making way for confirmation biases and invalid research. The problem of inductivism is marvelously exemplified by Bertrand Russel in Alan Chalmers book ‘What is this thing called science’[84]. He writes “*It concerns a turkey who noted on his first morning at the turkey farm that he was fed at 9 am. After this experience had been repeated daily for several weeks the turkey felt safe in drawing the conclusion ‘I am always fed at 9am’. Alas, this conclusion was shown to be false in no uncertain manner when, on Christmas Eve, instead of being fed, the turkey’s throat was cut. The turkey’s argument led it from a number of true observations to a false conclusion, clearly indicating the invalidity of the argument from a logical point of view.*” A number of true observations of animals begin sensitive to grammar produced stimuli leads the field to make the conclusion that it is the

grammar that is the drive of the behavior. The difference between the turkey and the AGL field is that the turkey does not have any have any other source of its data than when the farmer is feeding it, while the AGL field has a number of alternative explanations. However, these explanations are in the animal AGL field often not explicitly considered, and the underlying assumption of rule-learning results in experiments confirming hypothesizes instead of disproving them. It is clear that the work presented in this thesis is also not able to prove what the underlying mechanisms of AGL are. It might be that rule-learning – in the sense that animals learn to encode abstract rules – is the correct explanation for the phenomena. However, taken together, the deeply rooted assumption of rule-learning, the problems associated with confirmation/ inductive science, the lack of explicit controls for other possible explanations, and the an insufficient distinction between the behavior and its underlying mechanisms, I have to conclude that the current evidence for rule-based explanations of animal AGL is weak.

In this context it is interesting to draw a philosophical parallel with methods and behavior found a hundred years ago in a case of the horse Clever Hans. Clever Hans was a horse claimed to have the capacity to perform complex tasks such as arithmetic, telling the time and reading German. Obviously, the horse did not have these abilities but relied solely on unintentional cues from his trainer. If any researcher had been claiming any such capacities today, they would have been ridiculed. However, if one assumes first that the animal has this capacity, and then view the behavioral data in the light of that assumption, and do not control for other possible mechanisms, it would clearly show that the horse is capable of doing these intellectual tasks. The change in this story occurred when the horse, and its owner, were tested for a number of other possible explanations for the horse's performance, showing that there were a simpler explanation for the phenomena than what previously thought. The Clever Hans story shows that one needs to be cautious about drawing conclusions from what at first sight appears to be complex behavior, especially when there are simpler mechanisms or explanations available. Just because these simpler mechanisms or explanations could be harder to find and detail, it does not justify the leap to a complex conclusion. It shows that there need for better and more controlled animal AGL experiments that do not allow themselves to jump to conclusion on the basis that these conclusions would be interesting, or make an impact if they were true.

AGL as an experimental paradigm undoubtedly holds much power. But the AGL paradigm is a tool for investigation and experimentation, and as it is with any tool, it usually has only a few ways to be used correctly, but a million ways to be misused. In the same way, a screwdriver is primarily used for screwing screws, the number of potential wrong uses for a screwdriver is in the millions. By letting go of the tacit assumptions of rule-learning, and the narrow focus on grammar, one does not lose anything, but one opens for a wider view on animal cognition and the animals' capacities. For this reason, I encourage researchers using the AGL paradigm in animal cognition research to backtrack their chosen path of rule-learning and to reflect upon the number of issues that this thesis highlights. Identifying the paradigms correct uses with a critical mindset, conform to, or at least consider the

principles of Occam's razor, and to adhere to the frameworks brought forward in the philosophy of science would significantly improve the paradigms value, and would yield methods, experiments and theories with a much higher explanatory value.

In addition to improved explanatory power, there are other reasons to continue to use models in future investigations into animal AGL. One of these reasons is on a more ethical level. No responsible researcher wishes to cause any suffering or extensive amount of stress on their animals, and therefore it would be optimal to use as few animals as possible. Experiments on animals do however cause an arguably unjustified amount of stress on the animals involved. But using models to generate better and more rigorous hypotheses, before testing these in a laboratory, could maximize the potential value of each animal in the experiment. Modeling first, and then laboratory tests, could yield better and more accurate results, implying that fewer animals would be needed in experiments.

7 Conclusion

This chapter reviews the research questions posed in the introduction of this thesis, as well as evaluates the goals that were made. Then I discuss the novel contributions this research has given to the field of animal AGL research. Finally, I list the directions future research could take.

7.1. Project goals and research questions

This section described the research questions and their goals and answers them directly.

Research question 1: *What effect does the assumption of grammatical competence, in the form of rule-learning, in animal AGL experiments have on the conclusions these studies draw?*

To answer this research question, an analysis of the artificial grammar learning literature was completed. This review focused on an evaluation of *how* research is conducted, *why* it is conducted and *what* it concludes. By reviewing the literature, it is argued that animal research using the AGL paradigm faces two problems. Firstly, the design of the experiments are not very suitable to address the topic that the authors are interested in, namely to what extent animals have grammatical capabilities. This results in a biased explanatory narrative. Secondly, the animal AGL paradigm attempts to explain grammatical capabilities without considering the underlying mechanisms that the animal possess. These two problems result in the conclusion that the effect of having an underlying assumption of rule-learning when drawing conclusions from animal AGL experiments lead to a drop in explanatory power if the studies lack controls for other mechanisms.

Research question 2: *Can computational models of theoretical accounts of human AGL, other than rule-learning, account for the performance reported in animal AGL studies?*

To answer this question a reflection on the value of computational and cognitive models was given, before several models of word segmentation was presented. One of these models, PARSER, was used to replicate a number of animal AGL experiments. Two experiments that were replicated in silico show that the PARSER model is able account for the animals' reported performance. In a number of other experiments, it was tested if the model was able to learn a selection of string-sets from several animal AGL studies. This showed that the model is able to do so. From the results of these experiments it is concluded that computational models of theoretical accounts of human AGL can indeed account for the performance reported in animal AGL studies.

7.2. Contributions

This section summarizes the contributions this research project have made.

Contribution 1 *An alternative mechanism to rule-learning for explaining performance in animal AGL studies*

The animal AGL studies presented in this thesis report positive results in an AGL paradigm does so with the intention of exploring animal cognition, however, they have the tacit assumption that the animals are learning the rules of the grammar. This project has shown that there is at least one other cognitive mechanism – chunking – that can produce similar performance.

Contribution 2 *Modeling of animal cognition*

Computational and cognitive modeling as a means to investigate the emergence of behavior is a method that has previously almost exclusively been used in human and experimental computer science research. Computational models of human cognition, or at least models attempting to replicate human behavior have been around since the early dawn of the modern day computer. Yet modeling of animal behavior has only been used in experimental computer science and optimization algorithm research in attempts to either simulate collective animal behavior or solve various optimization problems. But these models rarely touch upon the inner workings of animal cognition in the same way models of human cognition does. The use of cognitive models to investigate animals' performance in AGL tasks is in of itself something new, but more generally it is novel by using models to examine and explain any part of animal cognition.

Another more general, but as important, contribution that comes with the introduction of modeling, is the addition of a framework for understanding complex systems such as cognition. This thesis introduced Marr's levels of analysis for understanding animal cognition. While Marr's framework is by no means a perfect framework, it is beneficial because it forces researchers to see beyond observed behavior and to analyze the possible mechanisms that result in the observed behavior. Such a framework improves the explanatory power of research not just by giving better explanations, but also by highlighting the limitations of providing an explanation that addresses only parts of the system. I believe that Marr's levels of analysis is a suitable framework of investigating complex systems such as animal cognition.

Contribution 3 *A python implementation of PARSER*

Models described in papers are often described in detail, but their implementations are rarely accessible. An important aspect of this research has been to encourage researchers to use models for hypothesis generation and to account for other mechanisms. Therefore, a Python implementation of PARSER is made publicly available for other researchers to use. The implementation of the model, and all the experiments conducted with is available at a GitHub repository[77].

7.3. Future research

The research here should be considered a primer for further research using cognitive models for investigating mechanisms of animal cognition. For this reason, there is a number of future directions this research could potentially take.

The first and probably most important, direction this research should take is to use models able to replicate animal AGL performance, for hypothesis generation. Simulating experiments using computational models can be used to generate a number of hypotheses about the animals' behavior in these experiments. These hypotheses can then, in turn, serve as guidelines for designing laboratory experiments. To use computational models for hypothesis testing has great potential as it creates an evaluation loop between the models and the real world data. The model improves the experiments, and the experiments can, in turn, be used to improve the models. This results in better explanatory power of the models and theories, better experiments, and overall a better and more grounded science.

A second path that should perhaps be traversed before my first suggestion, is the exploration of more models for performance modeling. The scope of this thesis was narrowed to a single word-segmentation model, so further research should broaden its scope to other word-segmentation models as well as models for describing other cognitive processes. SRNs, TRACX2 and other models shown to be able to account for AGL performance in human studies would be a good place to start this investigation. But other models, such as Long Short-Term Memory (LSTM) neural networks, have also shown great potential in word-segmentation and could yield potential value in the field of AGL as well. Hopefully, work on the model used in this paper, PARSER, would be continued as well. Further exploration of what stimulus PARSER is sensitive to, a mapping of PARSER's parameter space and an implementation of a grammatical inference method that uses the chunk-strengths from the vocabulary would greatly improve the model.

A final comment on what I encourage both modelers and animal researchers to do is to make their work accessible for others. The work of a modeler should be reproducible, and their implementation of their models should be available for others to use. And animal researchers should make their data available as well. Accessibility of data and models is an important step in using models to understand, explore and explain phenomena such as animal cognition.

8 References

[1] Y. Kikuchi, W. Sedley, T. D. Griffiths, and C. I. Petkov, "Evolutionarily conserved neural signatures involved in sequencing predictions and their relevance for language," *Curr. Opin. Behav. Sci.*, vol. 21, pp. 145–153, 2018.

[2] A. D. Friederici, J. Bahlmann, S. Heim, R. I. Schubotz, and A. Anwander, "The brain differentiates human and non-human grammars: functional localization and structural connectivity," *Proc. Natl. Acad. Sci.*, vol. 103, no. 7, pp. 2458–2463, 2006.

[3] R. C. Berwick, A. D. Friederici, N. Chomsky, and J. J. Bolhuis, "Evolution, brain, and the nature of language," *Trends Cogn. Sci.*, vol. 17, no. 2, pp. 89–98, 2013.

[4] G. J. L. Beckers, R. C. Berwick, K. Okanoya, and J. J. Bolhuis, "What do animals learn in artificial grammar studies?," *Neurosci. Biobehav. Rev.*, 2016.

[5] M. J. Spierings and C. ten Cate, "Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment," *Proc. Natl. Acad. Sci.*, vol. 113, no. 27, pp. E3977–E3984, 2016.

[6] B. Wilson, W. D. Marslen-Wilson, and C. I. Petkov, "Conserved sequence processing in primate frontal cortex," *Trends Neurosci.*, vol. 40, no. 2, pp. 72–82, 2017.

[7] T. Q. Gentner, K. M. Fenn, D. Margoliash, and H. C. Nusbaum, "Recursive syntactic pattern learning by songbirds," *Nature*, vol. 440, no. 7088, p. 1204, 2006.

[8] G. A. Miller, "Free recall of redundant strings of letters," *J. Exp. Psychol.*, vol. 56, no. 6, p. 485, 1958.

[9] A. S. Reber, "Implicit learning of artificial grammars," *J. Verbal Learning Verbal Behav.*, vol. 6, no. 6, pp. 855–863, 1967.

[10] B. J. Knowlton and L. R. Squire, "The information acquired during artificial grammar learning," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 20, no. 1, p. 79, 1994.

[11] P. Perruchet and C. Pacteau, "Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge?," *J. Exp. Psychol. Gen.*, vol. 119, no. 3, p. 264, 1990.

[12] E. M. Pothos, "Theories of artificial grammar learning," *Psychol. Bull.*, vol. 133, no. 2, p. 227, 2007.

[13] R. L. Gomez and L. Gerken, "Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge," *Cognition*, vol. 70, no. 2, pp. 109–135, 1999.

[14] G. F. Marcus, S. Vijayan, S. B. Rao, and P. M. Vishton, "Rule learning by seven-month-old infants," *Science (80-.).*, vol. 283, no. 5398, pp. 77–80, 1999.

[15] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *Science (80-.).*, vol. 274, no. 5294, pp. 1926–1928, 1996.

[16] G. Westphal-Fitch, B. Giustolisi, C. Cecchetto, J. S. Martin, and W. T. Fitch, "Artificial Grammar Learning Capabilities in an Abstract Visual Task Match Requirements for Linguistic Syntax," *Front. Psychol.*, vol. 9, 2018.

[17] N. Stobbe, G. Westphal-Fitch, U. Aust, and W. T. Fitch, "Visual artificial grammar learning: comparative research on humans, kea (Nestor notabilis) and pigeons (Columba livia)," *Phil. Trans. R. Soc. B*, vol. 367, no. 1598, pp. 1995–2006, 2012.

[18] C. M. Conway and M. H. Christiansen, "Modality-constrained statistical learning of tactile, visual, and auditory sequences.," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 31, no. 1, p. 24, 2005.

[19] N. Chomsky and G. A. Miller, “Finite state languages,” *Inf. Control*, vol. 1, no. 2, pp. 91–112, 1958.

[20] K.-M. Petersson, V. Folia, and P. Hagoort, “What artificial grammar learning reveals about the neurobiology of syntax,” *Brain Lang.*, vol. 120, no. 2, pp. 83–95, 2012.

[21] A. S. Reber and R. Allen, “Analogic and abstraction strategies in synthetic grammar learning: A functionalist interpretation,” *Cognition*, vol. 6, no. 3, pp. 189–221, 1978.

[22] G. J. L. Beckers, J. J. Bolhuis, K. Okanoya, and R. C. Berwick, “Birdsong neurolinguistics: Songbird context-free grammar claim is premature,” *Neuroreport*, vol. 23, no. 3, pp. 139–145, 2012.

[23] A. Attaheri, Y. Kikuchi, A. E. Milne, B. Wilson, K. Alter, and C. I. Petkov, “EEG potentials associated with artificial grammar learning in the primate brain,” *Brain Lang.*, vol. 148, pp. 74–80, 2015.

[24] J. Saffran, M. Hauser, R. Seibel, J. Kapfhamer, F. Tsao, and F. Cushman, “Grammatical pattern learning by human infants and cotton-top tamarin monkeys,” *Cognition*, vol. 107, no. 2, pp. 479–500, 2008.

[25] L. J. Rips, “Similarity, typicality, and categorization,” *Similarity Analog. Reason.*, vol. 2159, 1989.

[26] G. A. Miller, “Project Grammarama,” in *The psychology of communication*, G. A. Miller, Ed. New York, NY: Basic Books, 1967.

[27] W. T. Fitch, A. D. Friederici, and P. Hagoort, “Pattern perception and computational complexity: introduction to the special issue.” The Royal Society, 2012.

[28] C. A. Seger, “Implicit learning,” *Psychol. Bull.*, vol. 115, no. 2, p. 163, 1994.

[29] R. DeKeyser, “11 Implicit and Explicit Learning,” *Handb. Second Lang. Acquis.*, vol. 27, p. 313, 2008.

[30] H. L. Roediger, “Implicit memory: Retention without remembering,” *Am. Psychol.*, vol. 45, no. 9, p. 1043, 1990.

[31] A. S. Reber, “Transfer of syntactic structure in synthetic languages,” *J. Exp. Psychol.*, vol. 81, no. 1, p. 115, 1969.

[32] T. Johnstone and D. R. Shanks, “Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and Van der Linden (1997).,” 1999.

[33] D. E. Dulany, R. A. Carlson, and G. I. Dewey, “A case of syntactical learning and judgment: How conscious and how abstract?,” *J. Exp. Psychol. Gen.*, vol. 113, no. 4, p. 541, 1984.

[34] B. J. Knowlton and L. R. Squire, “Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 22, no. 1, p. 169, 1996.

[35] E. Servan-Schreiber, “The competitive chunking theory: Models of perception, learning, and memory.,” 1992.

[36] E. Servan-Schreiber and J. R. Anderson, “Learning artificial grammars with competitive chunking.,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 16, no. 4, p. 592, 1990.

[37] A. Cleeremans and J. L. McClelland, “Learning the structure of event sequences.,” *J. Exp. Psychol. Gen.*, vol. 120, no. 3, p. 235, 1991.

[38] A. Cleeremans, *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT press, 1993.

[39] P. Perruchet and S. Pacton, “Implicit learning and statistical learning: One phenomenon, two approaches,” *Trends Cogn. Sci.*, vol. 10, no. 5, pp. 233–238, 2006.

[40] Z. Dienes, G. Altmann, L. Kwan, and A. Goode, “Unconscious knowledge of artificial grammars is applied strategically,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 21, no. 5, p. 1322, 1995.

[41] J. Chen, N. Jansen, and C. Ten Cate, “Zebra finches are able to learn affixation-like patterns,” *Anim. Cogn.*, vol. 19, no. 1, pp. 65–73, 2016.

[42] S. A. Kotz, “A critical review of ERP and fMRI evidence on L2 syntactic processing,” *Brain Lang.*, vol. 109, no. 2–3, pp. 68–74, 2009.

[43] M. D. Hauser and W. T. Fitch, “What are the uniquely human components of the language faculty?,” *Stud. Evol. Lang.*, vol. 3, pp. 158–181, 2003.

[44] C. ten Cate, “Assessing the uniqueness of language: Animal grammatical abilities take center stage,” *Psychon. Bull. Rev.*, vol. 24, no. 1, pp. 91–96, 2017.

[45] R. A. Murphy, E. Mondragón, and V. A. Murphy, “Rule learning by rats,” *Science (80-.).*, vol. 319, no. 5871, pp. 1849–1851, 2008.

[46] J. M. Toro and J. B. Trobalón, “Statistical computations over a speech stream in a rodent,” *Percept. Psychophys.*, vol. 67, no. 5, pp. 867–875, 2005.

[47] M. C. Corballis, “Do rats learn rules?,” 2009.

[48] M. D. Hauser, N. Chomsky, and W. T. Fitch, “The faculty of language: What is it, who has it, and how did it evolve?,” *Science (80-.).*, vol. 298, no. 5598, pp. 1569–1579, 2002.

[49] M. C. Corballis, “Recursion, language, and starlings,” *Cogn. Sci.*, vol. 31, no. 4, pp. 697–704, 2007.

[50] Y. Suhara and A. Sakurai, “A simple computational model for classifying small string sets,” in *International Congress Series*, 2007, vol. 1301, pp. 270–273.

[51] C. A. A. Van Heijningen, J. De Visser, W. Zuidema, and C. Ten Cate, “Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 48, pp. 20538–20543, 2009.

[52] K. Abe and D. Watanabe, “Songbirds possess the spontaneous ability to discriminate syntactic rules,” *Nat. Neurosci.*, vol. 14, p. 1067, 2011.

[53] D. Marr, “A computational investigation into the human representation and processing of visual information,” *Free. San Fr. CA*, 1982.

[54] R. P. Cooper and D. Peebles, “Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanation,” *Top. Cogn. Sci.*, vol. 7, no. 2, pp. 243–258, 2015.

[55] P. Kitcher, “Marr’s computational theory of vision,” *Philos. Sci.*, vol. 55, no. 1, pp. 1–24, 1988.

[56] J. R. Saffran, E. L. Newport, and R. N. Aslin, “Word segmentation: The role of distributional cues,” *J. Mem. Lang.*, vol. 35, no. 4, pp. 606–621, 1996.

[57] M. R. Brent and T. A. Cartwright, “Distributional regularity and phonotactic constraints are useful for segmentation,” *Cognition*, vol. 61, pp. 93–125, 1996.

[58] M. H. Christiansen, J. Allen, and M. S. Seidenberg, “Learning to segment speech using multiple cues: A connectionist model,” *Lang. Cogn. Process.*, vol. 13, no. 2–3, pp. 221–268, 1998.

[59] P. Perruchet and A. Vinter, “PARSER: A model for word segmentation,” *J. Mem. Lang.*, vol. 39, no. 2, pp. 246–263, 1998.

[60] R. C. O’Reilly and Y. Munakata, *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press, 2000.

[61] M. C. Frank, S. Goldwater, T. L. Griffiths, and J. B. Tenenbaum, “Modeling human performance in statistical word segmentation,” *Cognition*, vol. 117, no. 2, pp. 107–125, 2010.

[62] J. R. Saffran, E. L. Newport, R. N. Aslin, R. A. Tunick, and S. Barrueco, “Incidental language learning: Listening (and learning) out of the corner of your ear,” *Psychol. Sci.*, vol. 8, no. 2, pp. 101–105, 1997.

[63] I. Giroux and A. Rey, “Lexical and sublexical units in speech perception,” *Cogn. Sci.*, vol. 33, no. 2, pp. 260–272, 2009.

[64] R. M. French, C. Addyman, and D. Mareschal, “TRACX: a recognition-based connectionist framework for sequence segmentation and chunk extraction.,” *Psychol. Rev.*, vol. 118, no. 4, p. 614, 2011.

[65] B. French and G. Cottrell, “TRACX 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2014, vol. 36, no. 36.

[66] R. M. French and D. Mareschal, “TRACX2: a RAAM-like autoencoder modeling graded chunking in infant visual-sequence learning.”

[67] N. Z. Kirkham, J. A. Slemmer, and S. P. Johnson, “Visual statistical learning in infancy: Evidence for a domain general learning mechanism,” *Cognition*, vol. 83, no. 2, pp. B35–B42, 2002.

[68] L. K. Slone and S. P. Johnson, “Statistical and chunking processes in infants’ and adults’ visual statistical learning,” in *Poster presented and the Biannual Conf. of the Society for Research in Child Development, April 2015, Philadelphia, USA*, 2015.

[69] J. L. Elman, “Finding structure in time,” *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[70] J. L. Elman, “Distributed representations, simple recurrent networks, and grammatical structure,” *Mach. Learn.*, vol. 7, no. 2–3, pp. 195–225, 1991.

[71] L. Boucher and Z. Dienes, “Two ways of learning associations,” *Cogn. Sci.*, vol. 27, no. 6, pp. 807–842, 2003.

[72] P. Cairns, R. Shillcock, N. Chater, and J. Levy, “Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation,” *Cogn. Psychol.*, vol. 33, no. 2, pp. 111–153, 1997.

[73] P. Perruchet and R. Peereman, “The exploitation of distributional information in syllable processing,” *J. Neurolinguistics*, vol. 17, no. 2–3, pp. 97–119, 2004.

[74] H. Terrace, “Chunking and serially organized behavior in pigeons, monkeys and humans,” *Avian Vis. Cogn.*, 2001.

[75] N. Fujii and A. M. Graybiel, “Representation of action sequence boundaries by macaque prefrontal cortical neurons,” *Science (80-.).*, vol. 301, no. 5637, pp. 1246–1249, 2003.

[76] X. Jiang, T. Long, W. Cao, J. Li, S. Dehaene, and L. Wang, “Production of supra-regular spatial sequences by macaque monkeys,” *Curr. Biol.*, 2018.

[77] Bror Emil Baklund Krogsrud, “PARSER-for-Python,” 2018. [Online]. Available: https://github.com/Bror-E/PARSER-for-Python.

[78] C. A. A. van Heijningen, J. Chen, I. van Laatum, B. van der Hulst, and C. ten Cate, “Rule learning by zebra finches in an artificial grammar learning task: which rule?,” *Anim. Cogn.*, vol. 16, no. 2, pp. 165–175, 2013.

[79] B. Wilson *et al.*, “Auditory sequence processing reveals evolutionarily conserved regions of frontal cortex in macaques and humans,” *Nat. Commun.*, vol. 6, p. 8901, 2015.

[80] B. Wilson, K. Smith, and C. I. Petkov, “Mixed-complexity artificial grammar learning in humans and macaque monkeys: evaluating learning strategies,” *Eur. J. Neurosci.*, vol. 41, no. 5, pp. 568–578, 2015.

[81] B. Wilson *et al.*, “Auditory artificial grammar learning in macaque and marmoset monkeys,” *J. Neurosci.*, vol. 33, no. 48, pp. 18825–18835, 2013.

[82] A. D. Endress, S. Carden, E. Versace, and M. D. Hauser, “The apes’ edge: positional learning in chimpanzees and humans,” *Anim. Cogn.*, vol. 13, no. 3, pp. 483–495, 2010.

[83] L. Chang and D. Y. Tsao, “The code for facial identity in the primate brain,” *Cell*, vol. 169, no. 6, pp. 1013–1028, 2017.

[84] A. F. Chalmers, *What is this thing called science?* Hackett Publishing, 2013.