

Universiteit
Leiden
The Netherlands

Simple rule learning is not simple : studies on infant and adult pattern perception and production

Geambasu, A.

Citation

Geambasu, A. (2018, December 11). *Simple rule learning is not simple : studies on infant and adult pattern perception and production*. *LOT dissertation series*. LOT, Utrecht. Retrieved from https://hdl.handle.net/1887/67540

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/67540

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle http://hdl.handle.net/1887/67540 holds various files of this Leiden University dissertation.

Author: Geambasu, A.
Title: Simple rule learning is not simple : studies on infant and adult pattern perception and production
Issue Date: 2018-12-11

Simple Rule Learning is Not Simple

Studies on Infant and Adult Pattern Perception and Production

The research reported here was supported by NWO Vrije Competitie grant 360.70.452 to Clara C. Levelt.

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6111
e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration: Copac (Tree), by Mircea Nechita

ISBN: 978-94-6093-310-3
NUR: 616

Copyright © 2018 Andreea Geambașu. All rights reserved.

Simple Rule Learning is Not Simple

Studies on Infant and Adult Pattern Perception and Production

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus Prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 11 december 2018
klokke 12.30 uur

door

Andreea Geambaşu

geboren 29 april 1986
te Boekarest, Roemënie

Promotores:	Prof. Dr. Clara C. Levelt Prof. Dr. Carel ten Cate
Promotiecommissie:	Prof. Dr. LouAnn Gerken (University of Arizona) Prof. Dr. Janet Grijzenhout Prof. Dr. Judith Rispens (University of Amsterdam) Prof. Dr. Niels Schiller

For my family,
past, present, and future,
perfect and imperfect

Contents

Acknowledgements ix

1 Introduction **1**
1.1 Artificial grammar learning 2
1.2 Domain specificity 3
1.3 Repetition 4
1.4 Familiarity 6
1.5 Ecological validity 7
1.5.1 Variability 7
1.5.2 Task and instruction 8
1.5.3 Grammar complexity 9
1.6 Summary 9

2 Rule learning in infants in the speech and general auditory domains **11**
2.1 Abstract 11
2.2 Introduction 12
2.2.1 The present work 14
2.3 Experiment 1 16
2.3.1 Methods 16
2.3.2 Results 19
2.3.3 Discussion 24
2.4 Experiment 2 24
2.4.1 Methods 24
2.4.2 Results 25
2.4.3 Discussion 27
2.5 Experiment 3 29
2.5.1 Methods 29
2.5.2 Results 31
2.5.3 Discussion 32

2.6 Combination of the 3 experiments . . . 32
2.6.1 Results . . . 33
2.7 General Discussion . . . 34
2.7.1 Repetition biases . . . 35
2.7.2 Future work . . . 36
2.8 Appendices . . . 37
2.8.1 Appendix A . . . 37
2.8.2 Appendix B . . . 38

3 Rule learning in infants in the visual domain 41
3.1 Abstract . . . 41
3.2 Introduction . . . 42
3.3 Method . . . 44
3.3.1 Participants . . . 44
3.3.2 Stimuli . . . 44
3.3.3 Procedure . . . 46
3.4 Results . . . 47
3.4.1 Data pre-processing . . . 47
3.4.2 Overall analysis . . . 48
3.4.3 Analysis per condition . . . 51
3.5 General Discussion . . . 54
3.5.1 Conclusion and future work . . . 55
3.6 Appendices . . . 57
3.6.1 Appendix A . . . 57
3.7 Acknowledgements . . . 57

4 Patterns in infant babbling 59
4.1 Abstract . . . 59
4.2 Introduction . . . 60
4.3 Language and Corpora Selection . . . 61
4.4 Corpus Data Collection . . . 62
4.5 Results . . . 63
4.5.1 Analysis . . . 65
4.6 General discussion . . . 70
4.7 Acknowledgements . . . 72

5 Rule learning in adults in the speech domain 73
5.1 Abstract . . . 73
5.2 Introduction . . . 74
5.3 Experiment 1 - Undirected familiarization . . . 77
5.3.1 Methods . . . 77
5.3.2 Results . . . 80
5.3.3 Reported strategies . . . 83
5.3.4 Discussion . . . 83
5.4 Experiment 2 - Directed familiarization . . . 83

5.4.1 Methods . . . 83
5.4.2 Results . . . 84
5.4.3 Discussion . . . 85
5.5 Experiment 3 - Implicitly directed testing . . . 85
5.5.1 Methods . . . 85
5.5.2 Results . . . 86
5.5.3 Discussion . . . 87
5.6 Experiment 4 - Feedback training . . . 87
5.6.1 Methods . . . 87
5.6.2 Results . . . 88
5.6.3 Discussion . . . 89
5.7 Overall analysis . . . 89
5.7.1 Stimulus variability . . . 89
5.7.2 Training length . . . 90
5.7.3 Reported strategies . . . 90
5.8 General Discussion . . . 91
5.8.1 Variation . . . 92
5.8.2 Feedback . . . 93
5.8.3 Instruction . . . 93
5.9 Conclusion . . . 94
5.10 Acknowledgements . . . 95

6 Rule learning with Lindenmayer grammars 97
6.1 Abstract . . . 97
6.2 Introduction . . . 98
6.3 Experiments 1 and 2 . . . 101
6.3.1 Methods and materials . . . 101
6.3.2 Descriptive statistics and results . . . 103
6.3.3 Discussion . . . 104
6.4 Experiment 3 . . . 107
6.4.1 Methods and materials . . . 107
6.4.2 Participants . . . 107
6.4.3 Stimuli . . . 107
6.4.4 Materials . . . 108
6.4.5 Procedure . . . 108
6.4.6 Analysis of the data . . . 109
6.4.7 Results . . . 109
6.4.8 Discussion . . . 113
6.5 General Discussion . . . 113
6.6 Acknowledgments . . . 115
6.7 Appendices . . . 115
6.7.1 Appendix A . . . 115

7 Conclusion **117**
7.1 Thesis summary 117
7.2 Domain specificity 119
7.3 Repetitions 120
7.4 Familiarity 122
7.5 Ecological validity 122
7.5.1 Variability 123
7.5.2 Task and instruction 123
7.5.3 Grammar complexity 125
7.6 Conclusion 126

References 127

References **127**

Summary 139

Samenvatting 143

Curriculum Vitæ 147

Acknowledgements

Thank you....

...to my supervisor and promotor Claartje Levelt for always encouraging me in my work, even when it felt like I wasn't achieving anything, for making me comfortable to share my thoughts and have an open dialogue, work related or otherwise. Thank you for the example you set of being a successfully and healthily balancing academic and non-academic life, and smiling through it all.

...to my project mates and dear friends Raquel Garrido Alhama and Michelle Spierings, two brilliant researchers with whom I've been lucky enough to be able to enjoy collaboration, exchanges of ideas, stimulating conferences, and perhaps most importantly, good meals and lots of laughs and hugs. I could not have dreamed up better peers in the project. I can only hope we can work together again in the future.

...to my co-promotor Carel ten Cate and to Jelle Zuidema for your feedback, direction, knowledge, and guidance throughout this project. And for the nice company as well.

...to my committee members for your work in reading and commenting on my thesis and for your feedback.

...to all my colleagues within the LUCL, the PhD Council, the LUCL Labs, and the LIBC for your support and discussions. Bobby Ruijgrok, Elly Dutton, Gulnaz Sibgatullina, Borana Lushaj, Daan van de Velde, Olga Kepinska, Viktorija Kostadinova, Anne van der Kant, Saskia Lensink, Gareth O'Neill, Yang Yang, Nurenzia Yannuar, Aliza Glasbergen-Plas, Lena Karvovskaya, Sima Zolfaghari, Kate Bellamy, Victoria Nyst, Leticia Pablos Robles. You all made this experience so much nicer.

...to my officemate for three years, Marieke Meelen, who made those first three years of this journey so bright and sunny.

...to Marijn van 't Veer, my colleague, then lab assistant, then officemate...thank you for your invaluable assistance in running some of the experiments in Chapter 2 as well as some work that didn't make it into this thesis, and for generally keeping the lab running smoothly.

...to Thijs Schama for programming the infant experiments, and to Jos Pacilly for building a beautiful new baby lab and for the invaluable support in getting that up and running. Thank you also to Monique Bisschop for your help as lab assistant during my first year.

...to Christina Bergmann and Helen Buckler who kindly welcomed me in Nijmegen to introduce me to the world of infant language acquisition research and methods.

...to my wonderful collaborators, local and international, who are co-authors in each relevant chapter but also deserve a mention here. To everyone who welcomed me so warmly to the MARCS Institute in Sydney, especially Paola Escudero, Samra Alispahic, Rosmin Dadwani, Varghese Peters. I learned a great deal. Andrea Ravignani, your enthusiasm is always contagious. Daan van Renswoude, Ingmar Visser, and Maartje Raijmakers, thank you for co-authoring a paper that may never have had a chance at publication without your work and perspective.

...to all the lovely the students who conducted their minor projects, internships, BA and MA theses with me and with Claartje, without whom this work would have taken at least twice as long: Astrid Gilein, Mirjam Bahlmann, Naomi Nota, Isabella Jordanoska, Renske Jacobs, Rabia Mahboeb, Johanne Rauwenhoff, Gosse Minnema, Floor Bonsel, Gijsbert Westland, Marjolein Maljaars, Lisa Kennedy, Yvette Spoolder, Sofja Volkova, and to Claire van Leeuwen, Mariska Scheel, and Laura Toron each of whom co-authored a chapter (paper).

...to the organizers and attendees of the Lorentz Workshop Comparative Biology of Language Learning (2017) for putting the wind back in my sails.

...to my paranymphs Marieke Meelen and Josefine Karlsson. I was lucky enough to have both of you in my life since day one of this journey, and I could not have done it without your emotional support, discussions, and laughter until the end.

...to my friends outside of academia who make sure I have something other than work to think and talk and laugh about.

...to my aunt and uncle, for your love.

...to my parents who have always encouraged and supported me in everything.

Finally, to Erdinç, thank you for all your love and patience and for being my light.

CHAPTER 1

Introduction

Human language is an acquired symbolic system of communication and thought. When acquiring language, (hearing) infants must learn a great deal, including the sound inventory of their language, how those sound categories may be combined in their native language to make words, and how the words may be combined into well-formed phrases. Along the way, they are also simultaneously learning the meanings and referents of these words and utterances. How infants are able to learn this complex system has been the focus of much research throughout the history of (psycho-)linguistics and developmental psychology going back centuries (W. Levelt, 2013). How *do* infants make the mental leap from their specific input to the ability to understand and produce an infinite number of utterances? What kinds of cognitive mechanisms are involved in learning the structures underlying language and are those mental mechanisms uniquely suited for learning language, or are they shared with other mental capacities? These are questions that will be addressed in this thesis using the tool of artificial grammar learning tasks.

The work presented here focuses on better understanding the biases infants and adults may have when learning rules, in both perception and production, whether these biases are domain-general or specific to language, how they may change in adulthood, and how they may interact with task-specific factors and attention. The experiments were developed as part of a larger project studying similar biases and learning mechanisms in birds (Spierings, 2016) and devising computational models of rule learning (Alhama, 2017).

The experimental work in this project was conducted by means of artificial grammar learning tasks. I will first introduce the motivation behind using

artificial grammar learning tasks, particularly with infants. A detailed outline of the thesis will follow, summarizing the biases and specificity of rule learning that will be addressed in each chapter.

1.1 Artificial grammar learning

At the most basic level, the task infants are facing when acquiring their native language involves discovering both discrete units, like phonemes and words, and the rules that regulate the ways in which these units can be combined. Despite what this highly simplified definition of language acquisition may suggest, it is a daunting task. Nevertheless, infants are highly sensitive to linguistic input, and already begin to perceive and process aspects of speech and language in the womb (Moon, Cooper, & Fifer, 1993). Speech perception experiments have revealed that soon after birth, infants already have the prosodic knowledge to discriminate their mother's language from languages with a different prosodic structure (Christophe & Morton, 1998), and can soon use statistical information in language to discover discrete units such as phoneme categories (Maye, Werker, & Gerken, 2002). How humans, particularly young infants, are subsequently able to form larger combinatorial units, such as words and sentences, from these discrete items is one of the main questions of linguistic research.

Over the past fifty years, beginning with work from Reber (1967), artificial language learning (ALL) or artificial grammar learning (AGL), paradigms have been important tools for studying our ability to extract discrete units from the speech stream, and to discover the rules combining them. The grammars used in AGL tasks may vary in their complexity and naturalness, but both the discrete units of which they are composed, and their structure, are typically selected such that potential confounds that can lead to solving the task in ways unrelated to the research question – which may appear in natural languages – are minimized. As such, AGL paradigms have been crucial in discovering learning abilities and limitations on learning across age groups, across species, and across domains (e.g., Knowlton & Squire, 1996; Perruchet & Pacteau, 1990; Gómez, 1997).

Following a seminal study from Saffran, Aslin, and Newport (1996) showing that eight-month-old infants are able to calculate statistical regularities in a continuous speech stream to extract words, the first AGL experiment studying infants' ability to learn and generalize rules to novel instances soon followed (Marcus, Vijayan, Bandi Rao, & Vishton, 1999). Using the Headturn Preference Procedure (Kemler Nelson et al., 1995), they exposed seven-month-old infants to a two-minute stream of synthetic speech composed of syllables arranged in an XYX, XYY, or XXY exposure grammar. During a subsequent testing phase in which novel syllables either followed the exposure grammar or a grammar inconsistent with the exposure, the infants showed sensitivity to the difference between the two grammars by longer looking times for the inconsistent test stimuli. The authors thus concluded that in addition to a sta-

tistical learning mechanism for learning fixed units, infants have an algebraic rule learning mechanism, where "algebraic" refers to the ability to extrapolate X and Y variables from the familiarization stimuli, allowing for combinations of novel syllables as long as they obey the combinatorial rule as a whole.

Marcus et al.'s (1999) work was seminal in the field of infant AGL, paving the way for novel experiments using variations on this testing paradigm and using different types of stimuli. Subsequent work focused on how these algebraic representations are formed (e.g., Gerken, 2006; Gerken & Bollt, 2008; Gerken, 2010), and on whether this type of learning was domain-specific (as claimed in Marcus, Fernandes, & Johnson, 2007) or domain-general (as claimed in Saffran, Pollak, Seibel, & Shkolnik, 2007). However, nearly two decades after the original publication of Marcus et al.'s initial paper, our understanding of the ability to learn these abstract rules is still lacking, and basic questions about perceptual biases that might influence learning, the domain-specificity of the process, and factors that might facilitate or hinder the learning process are still not answered adequately. Many extensions on the Marcus et al. paradigm have appeared in the literature, offering almost as many different accounts of the type of stimuli that facilitate, or lead to, learning, e.g., speech (Gerken, 2006, 2010; Gerken, Dawson, Chatila, & Tenenbaum, 2015), familiar stimuli (Saffran et al., 2007), stimuli with a communicative function (Ferguson & Lew-Williams, 2016), redundant stimuli (Thiessen, 2012), multimodal stimuli (Frank, Slemmer, Marcus, & Johnson, 2009), surprising stimuli (Gerken et al., 2015). However, despite several notable extensions on the work, the original findings were never replicated exactly. In fact, the literature gives us reason to believe that exact replication of this work is quite difficult. Most concretely, a footnote in Gerken's (2006) paper states that replication with the same age group and same grammars failed. The fact that there are no successful exact replications of the original work motivated the work conducted in the chapter 2 of this thesis. This was initially aimed at an extension of the Marcus study, but when this failed, we attempted to replicate the study as closely as possible. When this replication attempt failed, the focus of the thesis shifted to the study of a variety of factors that may influence (primarily) Marcus-type rule-learning abilities in infants as well as adults. The specific factors and themes that will be examined in this thesis are presented below, and the chapters in which these factors feature will be indicated.

1.2 Domain specificity

One of the goals of this thesis is to understand whether the ability to learn and generalize simple patterns in linguistic structure is part of a domain general learning capacity. Several studies have tested infants with Marcus grammars composed of non-speech stimuli. For instance Marcus et al. (2007) expanded on the 1999 rule learning work in a follow-up study showing that 7.5-month-old infants exposed to either the XYY or XYX pattern in stimuli consisting of

tones, timbres, or animal sounds were not able to learn the rule and generalize it to novel input in the same domain. However, if the infants were exposed to sung syllables, they could generalize the rule to novel sung syllables. In addition, the infants could transfer a rule learned on the basis of speech stimuli to test stimuli consisting of non-linguistic tones, timbres, and animal sounds. These findings indicate that rule learning is facilitated by speech, and that rules learned with speech stimuli can be transferred to non-linguistic stimuli. The authors thus argue that there may be a special role for speech in the ability to learn algebraic rules.

Others, however, have shown evidence of rule learning in the visual domain as well, albeit with limitations. Johnson, Fernandes and Frank (2009) tested eight and 11-month-olds on visual stimuli consisting of sequentially-presented triads of shapes. Although some learning was possible in this domain, Johnson et al. found limitations on generalization depending on whether the familiarization grammar contained an immediate repetition at the beginning (XXY), at the end (XYY), or not at all (XYX). The eight-month-olds could not discriminate the grammars when trained on triads with an early repetition (XXY) or a non-adjacent repetition (XYX). This age group only succeeded when they were trained on triads with a late repetition (XYY) and tested with inconsistent XYX stimuli. Infants at the age of 11 months could learn from familiarization stimuli with both early and late repetitions (XXY and XYY), but could still not learn from XYX familiarization stimuli. The early learning bias for late repetition patterns in visual stimuli is also found in Rabagliati, Senghas, Johnson, and Marcus (2012), where the stimuli were videos of sign language signs. The authors found that infants at 7.5 months old were able to generalize XYY but not XXY rules to novel input. Saffran et al. (2007), however, showed that infants could learn from both XYY and XYX visual stimuli if familiarization triads consisted of pictures of dogs or cats.

While rule learning in infants thus seems to be possible in non-linguistic domains, there appear to be limitations that need to be further explored, relating to repetition and the role of stimulus familiarity.

1.3 Repetition

Perception

One of the primary factors that seem to play a role in rule learning is repetition. A perceptual sensitivity to the presence of repetition in stimuli has been found across age groups (e.g., Gervain, Macagno, Cogoi, Peña, & Mehler, 2008; Gómez & Gerken, 1999; Endress, Scholl, & Mehler, 2005), species (e.g., Chen, van Rossum, & ten Cate, 2015; van Heijningen, Chen, van Laatum, van der Hulst, & ten Cate, 2013), and domains (e.g., S. P. Johnson et al., 2009). Endress and colleagues (Endress & Bonatti, 2007; Endress, Nespor, & Mehler, 2009; Endress, 2013) have characterized this sensitivity as one of the principal

factors aiding in the acquisition of simple statistics and rules during acquisition, which they have labeled Perceptual or Memory Primitives (POMPs). Infants in both speech and visual learning experiments seem more able to learn patterns involving immediate repetition than other patterns. Above, we discussed the visual rule learning work of Johnson et al. where infants could only learn from familiarization stimuli in XYY (8-month-olds) and XXY (11-month-olds) repetition patterns, but not from a non-adjacent repetition pattern XYX. The sensitivity to repetition patterns appears to be present already at birth: in a study using Near-infrared spectroscopy (NIRS), newborns showed more brain oxygenation in response to stimuli organized in the XYY pattern than to stimuli organized in a control pattern XYZ, while they did not show such a differential activation when the pattern XYX was pitted against XYZ (Gervain et al., 2008). It thus seems that, although these Marcus-type grammars appear at first glance equally simple, there is a hierarchy of learnability with respect to these Marcus-type rules, with XYY being the simplest, followed by XXY and XYX.

Learning of rules that do not involve immediate repetition may require extra support. Frank et al. (2009) show that five month old infants can learn both XYX and XYY rules when they are presented with congruent speech and visual information. They could not generalize the rules when they were instantiated only in the speech or in the visual domains. In this case, not only is the visual rule learning aided by speech, but rule learning in speech at this very early age is made possible by the presence of congruent visual information. Although the repetition bias can be overcome given the right input and learning circumstances, it nevertheless remains a strong perceptual bias throughout life. Adults in Marcus grammar learning tasks (Endress, Dehaene-Lambertz, & Mehler, 2007; Chen et al., 2015) and in more complex rule learning paradigms fall back on this POMP to help them generalize to novel input (Endress et al., 2005). The role of repetition is explored in chapters 2 and 3 of this thesis.

Production

Because the evidence for a bias for repetitions in perception is robust, both in the work discussed above and in our own work presented in chapters 2 and 3 of this thesis, we wanted to examine to what extent repetition is also present in production. The role of production abilities in perception preferences, and vice versa, is often neglected, but recent work shows interactions between the two (e.g., DePaolis, Vihman, & Nakai, 2013; Majorano, Vihman, & DePaolis, 2014; ter Haar & Levelt, 2018). To this end, a corpus analysis on babbling data of infants in their first year of life was conducted. Infants begin babbling at around the age of six months. The leading theories of the development of infant babbling state that between this age and the production of their first real words, infants go through a period of development in which they are thought to proceed from mostly repetitive (reduplicated) to more varied (variegated) syl-

lable productions (Stark, 1980; Oller, Yeni-Komshian, Kavanaugh, & Ferguson, 1980; Roug, Landberg, & Lundberg, 1989). Others have identified the entirety of this period as one of reduplication (Koopmans-Van Beinum & van der Stelt, 1986). There seems to be a general consensus with respect to the fact that certain sounds are produced and combined earlier than others by infants, and with respect to a rough time-course of when that happens. However, the data on which these stages have been based have come from a few babies, most of whom were acquiring English, and large-scale corpus analyses to back up these proposed stages are missing (cf. Lipkind et al., 2013).

With our study of the corpus of infant speech available in the CHILDES database, presented in chapter 4, we investigate these developmental stages in longitudinal babbling data from infants from eight different languages. Our aims were both to frame our perceptual findings within the context of what is going on in infants' production during the same period, and to check whether the proposed stages were generalizable to a wider range of languages. Results from the analysis of two-, three-, and four-syllable utterances show asymmetries between patterns including and not including immediate repetitions, with patterns including non-adjacent repetitions (XYX) being very infrequent.

1.4 Familiarity

While it is evident that repetition is one of the key factors governing how and whether infants learn and generalize rules, we also know that infants are able to learn far more complex rules than only those containing immediate repetitions. Other factors should therefore also be at play. Some authors have hypothesized that speech is not special *per se*, but that infants learn rules best on the basis of what is familiar to them (Saffran et al., 2007), and that they invoke rules to make stimuli less "surprising" (Gerken, 2010; Gerken et al., 2015). But the term "familiarity" is not well-defined. At which level stimuli should be familiar to infants for them to be able to take advantage of familiarity and to generalize rules is an empirical question. Familiarity may refer to specific items and whether they have been heard or seen before (e.g., specific phonemes, syllables, tones, shapes, or pictures). It may also refer to a more linguistic familiarity; for example, with respect to picture stimuli, it can refer to whether the participant has a lexical representation for the picture.

Saffran et al. (2007) claimed that failures and asymmetries in generalization in the visual domain (e.g., S. P. Johnson et al., 2009) may have arisen because those experiments had used shapes as stimuli, which were unfamiliar to infants of that age. The authors argued that if infants are familiar with the visual stimuli making up the Marcus-type rules, they should be able to learn them. To that end, they showed that seven-month-old infants were indeed able to generalize an XYX or XYY rule when they were familiarized with stimuli consisting of images of different dog or cat breeds (Ferguson & Waxman, 2015 showed similar results with infants between three and four months of age).

Saffran et al. (2007) argue that the success of the infants in this study can be attributed to the fact that infants are familiar with dogs and cats at this young age, and that pictures thereof are therefore categorizable for them. Learning with shapes is difficult because young infants have typically not yet been exposed to shapes as such. According to Saffran et al. it is the familiarity of speech that facilitates rule learning with speech stimuli, not the fact that speech is in the linguistic domain.

Yet the question remains, at what level were the dogs and cats in Saffran et al. familiar? Since infants are indeed often familiar with dogs and cats from very early on, they might even have an early lexical label for them (Bergelson & Swingley, 2012). If this is the case, it might mean that the visual rule learning task in Saffran et al. was more linguistic than intended. We thus hypothesized that if the reason for success in the Saffran et al. study was the fact that the infants' lexical knowledge facilitated categorization of the stimuli, rule learning may still be specific to language.

Chapters 2 and 3 both address issues related to differences in learning abilities as a result of the familiarity of the stimuli. In chapter 2, where we focus on rule learning in Marcus grammars from auditory stimuli, we use both speech (familiar) and birdsong (unfamiliar) stimuli. In chapter 3 we attempt to disentangle familiarity at the item level, i.e., stimuli "being familiar" from familiarity at the lexical level, i.e., stimuli "being lexically specified," by comparing the performance of 12- to 14-month-olds on a visual rule learning task in three conditions. In one condition the visual training stimuli consisted of images of familiar objects that were also typically lexically familiar to 12- to 14-month-olds; in another condition the images were of familiar objects that were typically not lexically familiar to 12- to14-month-olds; in the third condition images of nonsense objects were used, which were thus unfamiliar at both levels.

1.5 Ecological validity

It is striking that humans (both infant and adult) do not readily learn and generalize the Marcus-type grammars under all circumstances, while they are able to passively learn more complex artificial grammars (e.g., adults can learn finite state grammars such as BROCANTO which mimic the syntactic rules and word class distributions of natural languages, Fitch & Friederici, 2012), and infants learn the grammars of their native language. Another important factor that plays a role in the learnability of an artificial grammar like the Marcus-type one thus seems to be its ecological validity, referring to the naturalness of the input and the learning environment provided to the participant. Ecological validity may be addressed on multiple levels. Here, specifically, we look at the amount of variability present in the input, the type of learning task provided, and finally, the complexity and naturalness of the grammar itself.

1.5.1 Variability

Natural language contains a great deal of variability at all levels of input. The role of variability in language learning has therefore been studied in experimental settings too. Rost and McMurray (2009), for example, showed that 14-month-olds' learning of minimal pair words was improved if the words were produced by multiple speakers. Gómez (2002) showed that both adults and infants learned a non-adjacent dependency rule more reliably when the familiarization stimuli contained a large number of different units in-between the two units constituting the non-adjacent dependency. Gerken (2006) found that while nine month olds who were presented with a subset of four different triads from the Marcus et al. (1999) stimuli could generalize, infants who were presented with either XX*di* or X*di*X triads did not generalize to novel stimuli. In this case, they learned the specific rule of XX*di* or X*di*X. Having a single item in the Y position (i.e., no variability) thus seems to impede the creation of a Y variable, necessary for rule generalization.

In this thesis, the role of variability in the learning of Marcus-type rules is addressed in two ways, namely by paying attention to the phonological variability of the training stimuli (in chapter 2), and by varying the amount of different familiarization triads (in chapter 5).

Regarding phonological variability, the syllables used in the triads in Marcus and colleagues' original study, as well as in later extensions (e.g., Gerken, 2006) showed very little phonological variability: the familiarization stimuli contained only two different vowels, both of which were front vowels. Three of the four familiarization consonants had a coronal place of articulation, while all test consonants were stops. Four of the 12 triads had identical vowels when syllables were combined (*wilili, wididi, lewewe, dewewe*). One triad had identical consonants when syllables were combined (*wiwewe*). While infants in this experiment did show evidence of learning and generalization, the stimuli are hardly representative of the type of variety of phonemes and syllables found in natural language; we tried to remedy this in our extensions of the Marcus et al. paradigm in chapter 2.

Regarding amount of variability at the triad level, Gerken (2010) showed that when infants were first familiarized with the specific XX*di* rule and subsequently familiarized with five extra trials, two of which followed the XX*di* sub-rule, and three of which contained different syllables in the Y position, they were able to generalize to novel test stimuli following an XXY rule. The presence of some variability in the Y position thus allowed infants to switch from a local to a broad (algebraic) rule. Notably, only three relevant examples were sufficient for allowing generalization to occur. Gerken's later work even showed that under specific circumstances, generalization from a single repeated triad is possible (Gerken et al., 2015). In chapter 5, we try to get more insight into the necessary amount of variability for learning, by presenting participants training stimuli consisting of either three repetitions of 15 different triads, or 15 repetitions of three different triads.

1.5.2 Task and instruction

The last two chapters of the thesis, chapter 5 and chapter 6, discuss experiments performed with adult participants.

There have been few examples investigating Marcus-type rule learning in adults. This may be because we can successfully test adults with much more complex input, as has been shown from a wealth of experiments (e.g., Fitch & Hauser, 2004; Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006; Reber, Kassin, Lewis, & Cantor, 1980), which are summarized in more detail in the introductions of chapters 5 and 6. While Marcus-type rule learning has been shown to be successful in adults (Christiansen, Conway, & Curtin, 2000; Endress et al., 2007; Sun, Hoshi-Shiba, Abla, & Okanoya, 2012; Chen et al., 2015), adult participants in this task seem to be highly susceptible to the way the task is presented to them. In Christiansen et al. (2000)'s successful replication, adults needed to be instructed that they were participating in a pattern recognition task; they were unsuccessful otherwise (Christiansen, personal communication). There is reason to believe that for adults, rule learning and generalization may hinge on how they are instructed about the type of task they have to perform, and on whether they receive feedback or not during the familiarization or test phase of the experiment, particularly when it comes to simple grammars. While more complex grammars have been found to be more easily learned implicitly (Reber, 1976), simpler grammars, in which a variety of different strategies can be applied, may require the participants' attention to be directed to relevant features more explicitly. In natural language acquisition (e.g., in second language learning tasks), we also know that adults typically struggle with learning grammar in an implicit manner. Chapter 5 of this thesis systematically explores the effects of both task instruction and the type of task on the rule learning abilities of adults.

1.5.3 Grammar complexity

Properties of the grammar itself form another aspect that can influence learning. As mentioned above, adults are able to learn more complex grammars than Marcus-style grammars. Therefore, at first glance, it is surprising that adults might only be able to learn these simple grammars under specific task and instruction conditions. On the other hand, Marcus-style algebraic rules, and particularly the XYX rule, hardly reflect naturally-occurring phenomena in natural languages (see also the general discussion in chapter 2).

Lindenmayer grammars, or L-systems (Lindenmayer, 1968) have been used to investigate adults' ability to generalize rules (Saddy, 2009; Shirley, 2014). These grammars were originally developed to describe the cell growth of algae (Lindenmayer, 1968; Lindenmayer & Rozenberg, 1972) and have been applied to other plant structures (Prusinkiewicz & Lindenmayer, 1990; Samal, Peterson, & Holliday, 1994). Their recursive rewrite rules make them suitable for studying language processing as well, as their output is too complex to learn

by memorization of surface forms, and their underlying hierarchical rules parallel structures found in human language. For this reason, Saddy (2009) and Shirley (2014) began to explore these grammars with promising results. While adults in Shirley's work could discriminate the L-systems from non-L-system foils during testing, it was not clear on what basis this discrimination was taking place; the L-systems in the experiments were composed of two syllables, *bi* and *ba*, and according to the authors the surface manifestations of these grammars produced a rhythmic pattern that participants may have used to discriminate the L-system grammar from a foil grammar. In chapter 6, we extend on this work and attempt to disentangle the roles of task (continuing in the line of chapter 5), grammar complexity, and rhythm. Using rhythmic sounds instead of syllables, we touch on the issue of domain generality of rule learning (brought up in chapters 2 and 3 with infants), and we explore whether Lindenmayer grammars, being closer to natural language grammars, could be a promising alternative artificial grammar to the Marcus-style grammar for future rule learning studies.

1.6 Summary

With few exceptions (Dawson & Gerken, 2009; Johnson et al., 2009), empirical work on rule learning has been geared towards showing what infants can do, rather than providing a detailed pattern of successes and failures across ages. (Frank & Tenenbaum, 2011)

This thesis begins by tackling the problem identified in the above quotation, exploring in depth failures, as well as a few successes, in rule learning in several age groups and domains. In the end, I will bring together the findings of the five experimental chapters outlined above and discuss findings relevant to the theory of rule learning as well as observations related to methodological considerations.

CHAPTER 2

Rule learning in infants in the speech and general auditory domains

This chapter has been submitted and is under revision as Geambaşu, A., van Renswoude, D., Visser, I., Raijmakers, M., & Levelt, C.C. (under revision). Marcus et al. (1999) revisited: which mechanism underlies infants' abstraction of algebraic rules?

The chapter was the first attempt within an interdisciplinary project to establish a baseline for rule learning in young infants in the speech and the general auditory domains.

2.1 Abstract

In this study we attempted to extend on, and then replicate closely, a seminal study from Marcus et al. (1999) showing that seven-month-old infants use algebraic rules to generalize from their input to novel instances. In Experiment 1, we investigated whether infants were able to learn an XYX or XYY pattern and generalize it to different levels of abstractness: (a) familiar syllables in novel combinations, (b) familiar syllables previously heard only in Y position now heard in X position, and (c) completely novel syllables. In Experiment 2, we familiarized infants to the same patterns as in Experiment 1, but tested them only with novel syllables. In these first two experiments, we used a naturally-recorded, phonologically balanced set of phonemes as our stimuli. In Experiment 3, we reverted to a procedure and stimuli as similar as possible

to the original Marcus et al. (1999) study. Across the three experiments, using two different paradigms and using exposure and testing stimuli of gradually fewer degrees of difference from the original study, we were unable to replicate the results of Marcus et al. (1999). We show that indeed learning of these seemingly simple rules is, in fact, quite difficult and may be subject to specific constraints. Our findings showing a consistent preference for triads containing adjacent repetition line up squarely with the theory of perceptual or memory primitives (POMPs) of Endress et al. (2009), who have proposed that much of the artificial grammar learning findings can be explained through the recruitment of domain-general cognitive processes, including the sensitivity for edges and for repetitions (Mehler, Nespor, & Peña, 2008). This sensitivity has been found across species, domains, and age groups. Our results add support to the theory that repetitions, especially at edges, are highly salient for young learners.

2.2 Introduction

Learning and representing rules lies at the heart of understanding cognition and cognitive development. Rule learning is implicated in reasoning, in causal learning, in social learning, and in language learning (Bunge & Zelazo, 2006). Specifically in the domain of language and language acquisition, the ability to generalize combinatorial rules allows us to learn and apply properties of language structure beyond the input.

From a young age and without explicit instruction, infants are able to pick up on the regularities in their linguistic input related to frequencies, positions, and combinatorial possibilities of sounds within words, and words within phrases. Very young infants have been found to use statistical properties of language to learn about relevant categories and units. Infants between six and nine months of age have been shown to use the phonotactic patterns in their native language to recognize what is or is not a likely word in the language (Jusczyk & Luce, 1994). Between six and 7.5 months, they are also able to extract monosyllabic words from fluent speech (Jusczyk & Aslin, 1995), and at eight months, they can use transitional probabilities to extract multisyllabic words (Saffran et al., 1996).

Noticing regularities related to word segmentation seems to develop in parallel with the ability to learn how segmented units can be combined. Marcus et al. (1999) proposed that the ability to learn how items can be combined from a specific exposure set, and the ability to apply the extracted rule to novel items in another set, is done by learning abstract, "algebraic" rules, which make use of variables. They showed that already at seven months, infants were able to learn and generalize simple rules (XXY, XYY, XYX – henceforth referred to as Marcus rules) from a two-minute familiarization stream, where the relevant units were delimited by a pause (in Marcus et al., 1999 referred to as "sentences", here referred to as "triads"). This seminal study has been widely cited

as proof that infants have the ability to learn algebraic rules, quickly abstracting structure beyond their input.

Marcus and colleagues have since argued that speech is special when it comes to learning and generalizing such rules, as opposed to other domains, such as the visual domain (Frank et al., 2009), the general auditory domain (Marcus et al., 2007), or even the context of sign language (Rabagliati et al., 2012). Marcus et al. (2007) found that while 7.5-month-old infants were unable to learn to generalize the simple Marcus rules when they were formed of tones produced by musical instruments, they were able to apply such rules to the tone test stimuli when they had first been learned on the basis of speech stimuli. The authors concluded that speech might facilitate rule learning (or rule generalization) in domains where infants might otherwise not acquire rules. In a similar vein, Rabagliati et al. (2012) found that hearing infants at the same age were able to learn XYY, but not XXY, when the rules are composed of visual presentations of sign language signs. The authors concluded that the asymmetric results indicate that learning these patterns in a different domain is difficult and that speech may be special, at least to hearing children who are learning from speech (as opposed to non-hearing children for whom sign might be special). Frank et al. (2009) found that five-month-old infants could only learn an XYX or XYY pattern from multimodal, concordant exposure to visual and speech stimuli. Learning from either of the unimodal speech or visual conditions did not occur, indicating that even with speech, the mechanisms for generalization are not yet available to such young infants unless they are supported by redundant information.

Other research has countered the hypothesis that speech is special. Ferguson and Lew-Williams (2016) showed that seven-month-olds could learn XYY rules from tones when they were pre-exposed to a video where tones were used as communicative signals. In the visual domain, Saffran et al. (2007) found that seven-month-old infants were able to learn Marcus rules with images of either different dog breeds or different cat breeds, indicating that the ability is not domain specific. In this case infants succeeded in both an XYX vs. XYY or XYY vs. XXY condition. Ferguson and Waxman (2015) also found that infants as young as three months old could perform the same task, when the visual patterns were accompanied by child-directed speech (not following the same pattern). Adding to the mixed findings, S. P. Johnson et al. (2009) found that eight- and eleven-month-old infants were not always able to learn rules from visual stimuli consisting of shapes. The younger infants were only able to discriminate a non-adjacent repetition when trained on an late repetition (XYY vs. XYX) but not vice versa. The older infants could discriminate early repetitions (XXY) from late repetitions (XYY), but they could only discriminate XXY from XYX when familiarized with XXY. An XYX training grammar thus did not result in rule learning by either age group. From this work we get a more nuanced view that suggests that rule learning is possible in the visual domain but that it may be subject to more cross-developmental variability.

This work also suggests that rules composed of adjacent repetitions are learned more easily.

In the speech domain, Kovács and Endress (2014) showed a similar finding in a rule learning experiment with seven-month-olds in which various syllables were used to create XYX-, XYY-, and XXY-structured words that were subsequently organized hierarchically within either XYY- or XXY-structured sentences. They found that infants could learn the sentence patterns containing adjacent repetitions, but not those containing non-adjacent repetitions. Their work also shows that sensitivity to adjacent repetitions may be an important cue for generalization.

Despite these extensions on the original Marcus et al. (1999) paradigm, the only published replication using the same stimuli and paradigm is presented in Gerken (2006). However, this study still did not constitute a true replication and contains a number of differences with Marcus and colleagues' original work. Gerken could not replicate the original findings with seven-month-olds but was able to do so with nine-month-olds. In addition, the results of Marcus et al. (1999) could only be replicated when an early repetition grammar was used (XXY), instead of the original late repetition grammar (XYY); pilot testing contrasting XXY and XYY yielded no learning and a "marginal" preference for XYY. With respect to the stimuli, Gerken used only one block of four triads in the familiarization as opposed to the 16 in Marcus et al., and only four triads in the test as opposed to three blocks of the four triads as Marcus et al. did. A footnote reveals that using the full set of 16 familiarization triads failed to replicate (when testing XYY vs. XXY). Finally, unlike Marcus et al. , Gerken found a familiarity effect instead of novelty effect.

The mixed findings in experiments that vary in different ways from the original Marcus et al. (1999) experiment are not trivial, as they indicate that the original findings may not be robust or may only hold under specific conditions. The question remains of under what conditions infants are able to learn such rules and what types of cues they may be sensitive to.

2.2.1 The present work

In the present work, we attempted to extend on the Marcus et al. (1999) and the Marcus et al. (2007) findings to understand under which conditions Marcus rules are learnable. In Experiment 1, we aimed to test whether age, stimulus type, or specific information in the test trials would either support or hinder rule generalization. To this end we created two exposure and test conditions, with the intention of testing whether infants better learned the rules when they were composed of speech sounds or whether the natural, harmonic properties of birdsong would also support learning. These conditions will henceforth be referred to as the Speech and Song conditions.

In the Speech condition, we created natural recordings[1] of a number of syllables. We hypothesized that the type of learning shown in Marcus et al. (1999) should only be enhanced when using natural speech. We thus took care to balance the syllables' phonological features with the intention of being more representative of natural speech than Marcus et al.'s (1999) set, which uses a very limited set of sounds. Some examples of the limitations include the fact that familiarization triads were composed of only two vowels, /e/ and /i/, both of which are front vowels, that all test consonants are stops, that all familiarization syllables and three of the four test syllables are homorganic[2], that three of the four familiarization consonants are coronal, that four triads had identical vowels when syllables were combined (*wilili, wididi, lewewe, dewewe*), and that one triad had identical consonants when syllables were combined (*wiwewe*). Table 2.1 shows syllables used in Marcus et al. (1999) as compared to the syllables used in our experiments.

In the Song condition, we aimed to directly test the findings of Marcus et al. (2007) that speech is special by testing learning in a non-speech condition with auditory stimuli found in nature[3].

In both conditions, we tested infants with a variety of test trial types, including previously-heard syllables in novel combinations maintaining the original ordinal positions within triads (Combination), previously-heard syllables in novel positions (Place), and completely novel syllables (Generalization). With these three types of testing conditions, we aimed to draw comparisons between the abilities of infants, adults (Geambaşu, Spierings, ten Cate, & Levelt, in prep.), and birds (Spierings & ten Cate, 2016) within a larger comparative project. While zebra finches were able to apply the learned rule only to familiarized sounds (Combination), budgerigars were able to apply the learned rule at all levels (Spierings & ten Cate, 2016). We expected the infants in this experiment would also be able to do this: if algebraic learning has occurred, variables are formed for X and Y items and a rule is formed for how to combine them, irrespective of the familiarity and positions of syllables. However, this experiment failed to deliver evidence of learning. Due to the complex nature of the experiment, we simplified the test and honed in on Generalization items using only speech stimuli in Experiment 2. When this experiment also failed to show evidence of learning, we conducted Experiment 3, a failed attempt at a close replication of Marcus et al.'s original study. In next sections of this paper

[1]Since we used natural birdsong we decided to use natural speech stimuli as well. In addition, Saffran and Wilson (2003) pegged the high drop out rate in their experiment on the disinterest of the infants in synthetic synthesized speech. Thus, natural recordings were considered more likely to engage the infants in the task.

[2]Homorganic sounds are those that share place of articulation. For example, coronal consonant /d/, and coronal, close-mid, front vowel /e:/ share the feature of being produced with the coronal part of the tongue in the front of the mouth, making /de:/ a homorganic syllable. Infants produce homorganic syllables first (C. C. Levelt, 1994) and prefer these sounds over non-homorganic ones early in development (ter Haar & Levelt, 2018).

[3]Birdsongs were also used for the purpose of cross-species comparison work conducted by project colleagues (Spierings & ten Cate, 2016).

Table 2.1: Syllable chart of possible consonant and vowel combinations. Syllables boxed with double lines were used in Marcus et al. (1999)'s experiment. Syllables with in boxes with polka dot fill were used for as test syllables in Marcus et al. (1999). Familiarization and test syllables were varied for each participant in our experiments. Syllables in bold were used in our experiments 1 and 2. Syllables with boxes with single lines replaced syllables /dʒi/ and /dʒe/ in our Experiment 3. Homorganic syllables used in the experiments are underlined. Unused syllables are grayed out.

Consonants / Vowels		/a:/	/e:/	/i/	/o:/	/u/
Voiced	Features	dorsal, back, open	coronal, front, close-mid	coronal, front, close	labial/round, back, close-mid	labial/round, back, close
/m/	labial	ma:	me:	**mi**	**mo:**	mu
/b/	labial	ba:	be:	**bi**	**bo:**	bu
/v/	labial	va:	ve:	vi	vo:	vu
/w/	labial	wa:	we:	wi	wo:	wu
/dʒ/	coronal	dʒa:	dʒe:	dʒi	dʒo:	dʒu
/d/	coronal	da:	**de:**	di	**do:**	du
/l/	coronal	la:	**le:**	li	lo:	**lu**
/n/	coronal	na:	***ne:***	**ni**	**no:**	**nu**
/g/	dorsal	ga:	ge:	gi	go:	gu
Unvoiced						
/f/	labial	fa:	fe:	**fi**	**fo:**	fu
/p/	labial	pa:	pe:	pi	po:	**pu**
/s/	coronal	**sa:**	se:	si	so:	**su**
/t/	coronal	**ta:**	***te:***	**ti**	to:	tu
/k/	dorsal	**ka:**	**ke:**	ki	ko:	ku
/X/	dorsal	ga:	**ge:**	gi	**go:**	***gu***

we describe the three experiments in detail and present both a frequentist and Bayesian analysis of the data.

2.3 Experiment 1

In Experiment 1, we extended on the foundation of Marcus et al.'s (1999) study. We exposed two age groups (six- and nine-month-old infants) to one of two patterns (XYX or XXY), in a Song or a Speech condition, and tested their ability to apply the learned rule to various levels of novelty during test.

2.3.1 Methods

Stimuli

Speech condition. Because the stimuli used in the original Marcus et al. (1999) study were phonologically limited, in the Speech condition, we expanded the group of syllables used as exposure and test stimuli. We balanced the number of syllables that used voiced and unvoiced consonants, and the number of syllables composed of homorganic and not homorganic sounds. In addition, our triad combinations differed from Marcus et al.'s in that all syllables could be used as either an X or Y item, as long as triad followed the rule.

The stimuli in the Speech condition consisted of natural recordings of consonant-vowel (CV) syllables made by a phonologically trained female native speaker of Standard Dutch (C.C.L.). The original CV recordings were manipulated such that each syllable was approximately 400 ms in total duration, all vowels were similar in length, and all consonants of the same class (stops, fricatives) were similar in length (vowel duration values taken from Koopmans-van Beinum, 1980). A total of eight unique syllables were selected[4], and each syllable was assigned to a letter A through H, in eight different randomizations (see Appendix A1 for a table of all syllables and to which letter they corresponded to in each randomization). These randomizations were made such that infants' performances could not be subject to biases or preferences for specific syllables. The syllables were combined into triads by concatenating three CV syllables as shown in Table 2.2 (see also Appendices B1 and B2 for a complete list of triads per randomization for the familiarization and test phases respectively). The CV syllables within a triad were separated by 100 ms silent pause. This created triads approximately 1.4 seconds in duration. Both the familiarization and test strings were concatenated with a one second pause separating each triad. Each triad's pitch was standardized at a constant 125 Hz such that pitch sounded natural for a female voice and would not highlight any specific syllable in the triad, thus providing no supporting information with respect to the structure of the triad.

[4]Eight syllables could be selected so as to match the number of unique zebra finch song elements available.

Table 2.2: Each letter A through H was assigned to a unique syllable. Syllables were then combined to form 30 familiarization triads and 18 test triads, half of which followed the XYX pattern and half of which followed the XXY pattern. To avoid influence of preference for specific sounds, each infant heard one of the possible eight combinations (see Appendices A1, B1, and B2).

Familiarization triads		**Combination test triads**		**Position test triads**	**Position test triads**	**Generalization test triads**	
XYX	**XXY**	**XYX**	**XXY**	**XYX**	**XXY**	**XYX**	**XXY**
ADA	AAD	ACA	AAC	FAF	FFA	GHG	GGH
AEA	AAE	BDB	BBD	FEF	FFE	HGH	HHG
AFA	AAF	CEC	CCE				
BAB	BBA	DCD	DDC				
BCB	BBC	EAE	EEA				
BFB	BBF						
CAC	CCA						
CDC	CCD						
CFC	CCF						
DBD	DDB						
DED	DDE						
DFD	DDF						
EBE	EEB						
ECE	EEC						
EFE	EEF						

Song condition. Stimuli used in the Song condition were also composed of natural recordings of eight zebra finch song elements produced in the Leiden University Institute of Biology. The duration of each element was between approximately 35 and 86 ms. As in the speech condition, triads were concatenated with 100 ms between each element, and strings were concatenated with a one second silent pause between triads. Also as in the Speech condition, all elements could be used as either an X or a Y, as long as the triad followed the rule. Song triads ranged in duration from approximately 400 ms to 560 ms including pauses.

Participants

A total of 19 six-month-olds and 20 nine-month-olds were tested in the Speech condition; 29 six-month-olds and 27 nine-month-olds were tested in the Song condition.

Infants were recruited via the municipality of Leiden. Parents were briefed on the intention of the study and subsequently gave their informed consent for their infants' participation. Parents and infants were compensated for their voluntary participation with reimbursed travel costs and a gift of a book.

Infants were removed from analysis for the following reasons: experimenter or technical errors (n=6), birth more than three weeks pre-term[5] (n=7), and excessive fussiness (n=4).

Sixteen infants were bilingual to some extent, ranging from hearing some foreign language from a grandparent to having at least one parent who spoke a foreign language to them regularly. Most of the bilingual infants only participated in the Song condition such that the speech stimuli would not be more foreign to them than for the rest of the subject group. The three bilingual infants who were tested in the speech condition had Dutch mothers and heard Dutch most of the time.

While Marcus et al. (1999) tested seven-month-old infants, we tested six- and nine-month-olds. In testing on either side of this age, we expect to see a developmental difference in generalization abilities.

Procedure

Infants were tested in the Visual Fixation Procedure (VFP; Cooper & Aslin, 1990) using E-Prime software (*E-Prime 2.0, Psychology Software Tools, Inc.*, 2012). While the VFP testing method is different from the Headturn Preference Procedure (HPP) used in Marcus et al. (1999), both methods rely on the same principle that infants will continue to look towards a visual stimulus for as long as they are attending to the auditory stimulus playing at the same time (E. K. Johnson & Zamuner, 2010).

The experiment consisted of an initial attention grabber, a familiarization phase, and a testing phase. The experiment began with the infants being seated

[5]According to the protocol of the Many Babies Consortium (unpublished protocol).

on a caregiver's lap in a sound-attenuated booth. Caregivers wore tight-fitting headphones that played a mixture of music and backwards speech to mask the auditory stimuli of the experiment. This ensured that caregivers would not reflexively respond to the auditory stimuli and unwittingly influence the behavior of the infant. The experiment began with an initial attention grabber, composed of a video of green LED lights, to draw the infant's attention to the center. Once the infant looked towards the attention grabber for two continuous seconds, the familiarization phase would begin. The familiarization phase consisted of an approximately two-minute (1 minute 53 seconds) continuous exposure phase. The continuous auditory stream included three repetitions of 15 different triads. While the auditory stimuli played, infants saw a video of a flashing red light on the screen directly in front of them to maintain their attention. Once the familiarization stream ended, the attention grabber played again. Once the infant looked towards the attention grabber for two seconds, the testing phase began. During the testing phase, infants saw the same video of flashing lights as during the familiarization. However, in contrast to the familiarization phase, the test trials only played while the infant was looking. Infants could look to a trial for a maximum time of 15 seconds, after which the trial would automatically end. If the infant looked away during the trial for more than two seconds continuously, the trial would automatically stop and an attention grabber would begin until the infant again focused to center for two continuous seconds.

During the test phase, infants heard three different types of test triads. The first, we refer to as Combination triads, which were made up of syllables infants had heard during familiarization but which were now rearranged into previously unheard combinations. The second, we refer to as Place triads, composed of syllables in which the X test item had only been heard in the Y position during the familiarization. The third type of test type we refer to as Generalization triads. These were, as in the original Marcus et al. (1999) experiment, triads composed of syllables which had not been heard during the familiarization phase. Half of all test triads followed the pattern of the familiarization phase (consistent test items), and half followed the opposite pattern (inconsistent test items). The test items were presented in an unblocked manner and randomized per participant. The test phase included 18 total test trials (see Table 2.2 and Appendix B2) resulting in a test phase of approximately 10 to 15 minutes.

A video camera recorded the infants behavior for offline coding and allowed the experimenter to react to the infants' behavior online. Infants' behavior was coded offline by a coder blind to the stimuli using ELAN software (*ELAN Version 5.0.0*, 2017; Sloetjes & Wittenburg, 2008) and precise looking times were calculated for each trial.

2.3.2 Results

Data preprocessing

Apart from the inclusion and exclusion of outliers, there are many decision to be made regarding the preprocessing of infant looking data. For instance, it is common practice to exclude looking times (LTs) that are shorter than the time it takes for one repetition of the stimulus to play (Ferguson & Waxman, 2015; Saffran & Wilson, 2003). Following the latter model, the data is analyzed without LTs shorter than 1.5 seconds for the speech and shorter than 0.5 seconds for the sound stimuli (10% of the data).

Another consideration is that there are LTs that exceed the 15 second presentation time of the stimuli (3% of the data). These LTs could be caused by infants not responding to the stimuli but simply staring (or, in the case of the Headturn Preference procedure, reflect a preference for one side of the booth). Ideally, only data points reflecting a preference would be kept and staring behavior would be omitted. However, this distinction cannot be made based on the data, and the literature provides no clear way to handle such cases, although removal of LTs reaching the maximum trial duration have been removed in previous studies (Bernard & Gervain, 2012; Gervain & Werker, 2013). We therefore decided to report the analyses both with and without LTs above 15 seconds.

In all experiments, data was first analyzed using repeated measures ANOVA (labeled 'Frequentist analysis' per experiment) and subsequently using Bayesian paired samples t-tests (labeled 'Bayesian analysis' per experiment). In addition to the more commonly-used frequentist analyses, we choose to include Bayesian analyses as they allow us to quantify relative evidence of competing hypotheses (Wagenmakers, Morey, & Lee, 2016). That is, a certain hypothesis (H1) can be more likely than another hypothesis (H0) or vice versa, given the data. In our studies H0 and H1 are clearly defined: "Infants will show preferential looking based on the test grammar's inconsistency with the familiarization grammar" (H1) or "Infants Infants will not show preferential looking based on the test grammar's inconsistency with the familiarization grammar" (H0). An additional benefit of this procedure is that null-results can be interpreted, since in a Bayesian analysis, H0 can be 'x' times more likely than H1 given the data. Such an interpretation is impossible with frequentist analyses, since the truth of the null-hypothesis is an assumption of frequentist analyses (Cohen, 1994).

Frequentist analysis

There were three conditions (Combination, Generalization, and Place) with two type's of stimuli (Song and Speech) creating six outcome options. Together this results in 12 LTs to compare between the Consistent and Inconsistent condition. To make the matter more complicated, there are also possible interactions with sex, age, and type of training stimuli (i.e., XYY or XXY) that should be

			Combination	Generalization	Place
LT > 15 allowed	Speech	Consistent	$F(1, 28) = 0.80, p = .37$	$F(1, 26) = 0.06, p = .80$	$F(1, 28) = 0.00, p = .96$
		Age:Consistent	$F(1, 28) = 1.72, p = .19$	$F(1, 26) = 0.01, p = .89$	$F(1, 28) = 0.15, p = .70$
		Sex:Consistent	$F(1, 28) = 0.33, p = .56$	$F(1, 26) = 0.06, p = .79$	$F(1, 28) = 0.19, p = .66$
		Training:Consistent	$F(1, 28) = 1.96, p = .17$	$F(1, 26) = 3.09, p = .09$	$F(1, 28) = 0.27, p = .60$
	Song	Consistent	$F(1, 44) = 0.12, p = .72$	$F(1, 42) = 0.26, p = .60$	$F(1, 42) = 4.50, p = .03$
		Age:Consistent	$F(1, 44) = 0.03, p = .86$	$F(1, 42) = 4.02, p = .05$	$F(1, 42) = 0.57, p = .45$
		Sex:Consistent	$F(1, 44) = 0.03, p = .84$	$F(1, 42) = 0.59, p = .44$	$F(1, 42) = 0.44, p = .51$
		Training:Consistent	$F(1, 44) = 0.17, p = .67$	$F(1, 42) = 5.88, p = .01$	$F(1, 42) = 0.70, p = .40$
LT > 15 omitted	Speech	Consistent	$F(1, 28) = 0.94, p = .33$	$F(1, 26) = 0.06, p = .80$	$F(1, 27) = 0.01, p = .89$
		Age:Consistent	$F(1, 28) = 1.49, p = .23$	$F(1, 26) = 0.01, p = .89$	$F(1, 27) = 0.22, p = .63$
		Sex:Consistent	$F(1, 28) = 0.42, p = .51$	$F(1, 26) = 0.06, p = .79$	$F(1, 27) = 0.32, p = .57$
		Training:Consistent	$F(1, 28) = 2.20, p = .14$	$F(1, 26) = 3.09, p = .09$	$F(1, 27) = 0.15, p = .69$
	Song	Consistent	$F(1, 44) = 0.19, p = .65$	$F(1, 41) = 0.68, p = .41$	$F(1, 42) = 2.24, p = .14$
		Age:Consistent	$F(1, 44) = 0.03, p = .84$	$F(1, 41) = 1.98, p = .16$	$F(1, 42) = 0.09, p = .75$
		Sex:Consistent	$F(1, 44) = 0.59, p = .44$	$F(1, 41) = 0.83, p = .36$	$F(1, 42) = 2.11, p = .15$
		Training:Consistent	$F(1, 44) = 0.66, p = .42$	$F(1, 41) = 6.05, p = .01$	$F(1, 42) = 0.13, p = .71$

Figure 2.1: Multiverse of p-values for Experiment 1.

assessed, creating 48 p-values to evaluate. This expansion is called the data multiverse (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016) and is very common in many psychological experiments, although it is almost never made explicit that there were so many outcome options. In order to increase transparency we decided to report all results. Figure 2.1 highlights p-values below .05 in gray and clearly shows an overall failure to replicate earlier findings (Gerken, 2006; Marcus et al., 1999). The only significant main effect (if we do not consider multiple comparison issues) occurs in the Place testing trials in the Song condition, and is in the opposite direction: LTs are longer during consistent than inconsistent trials. The significant interaction effects between training and condition are visualized in Figure 2.2 with error bars of ± 1 standard error. These effects indicate that infants have a preference for adjacent repetition (i.e., a preference for XXY versus XYX), independent of familiarization grammar.

Bayesian analysis

It is difficult to draw conclusions from the frequentist analysis presented above, because null results cannot be interpreted within the frequentist framework.

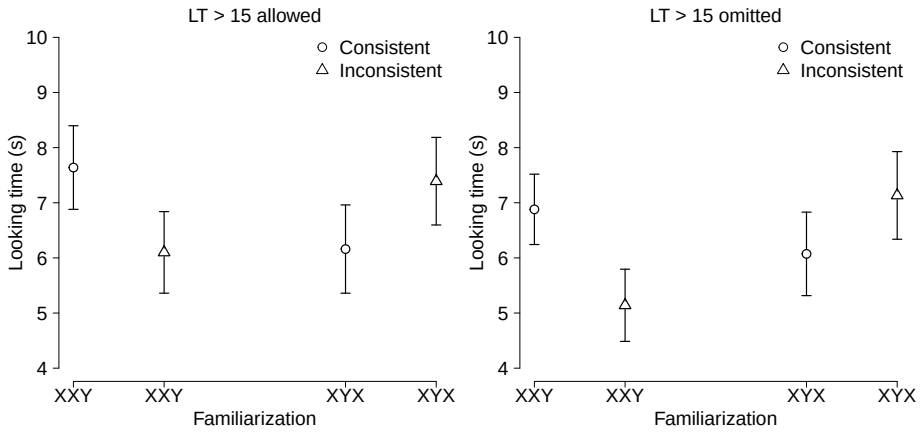


Figure 2.2: Interactions of familiarization and test grammars in Experiment 1. The left panel shows results when LTs of longer than 15 seconds are allowed in the analysis, while the right panel shows results when LTs of longer than 15 seconds are omitted from analysis. inclusions or omission did not change the pattern of results: infants familiarized with XXY look longer to the consistent test grammar, while infants familiarized with XYX look longer to the inconsistent test grammar.

Yet it would certainly help our understanding of infants' behavior if we could interpret these null effects. The Bayesian framework provides the possibility to accumulate evidence in favor of the null hypotheses and interpret how likely the null hypothesis is in comparison to other hypotheses (Jeffreys, 1961). This is possible by calculating Bayes Factors (BFs), which indicate how many times more likely one hypothesis is over another. In this case, we use BFs to quantify how much more likely the null hypothesis (that looking times are equal for both the consistent and inconsistent sounds) is than the alternative hypothesis (that infants look longer when they hear inconsistent sounds).

To make this comparison we use JASP (JASP Team, 2017), an alternative for SPSS using Bayesian analyses. In the Bayesian framework it is possible to accumulate evidence with every new data point. This can be seen in Figures 2.3 and 2.4 where the number of participants is plotted on the x-axis and the BF is plotted on the y-axis. With the data of every new participant, the BF is updated and the evidence accumulates. The sequential analyses for the Speech (Figure 2.3) and Song (Figure 2.4) conditions for the generalization test trials are shown, with LTs longer than 15 seconds allowed. Using BFs does not solve the problem that there are many possible comparisons that can be made. In Table 2.3 we report Bayes Factors that indicate how much more likely the null hypothesis is than the alternative hypothesis for three types of test trials (Combination, Generalization, Place), two stimulus conditions (Speech, Song), and two types of preprocessing options (with and without looking times longer than 15 seconds). Note that for every combination the null hypothesis is

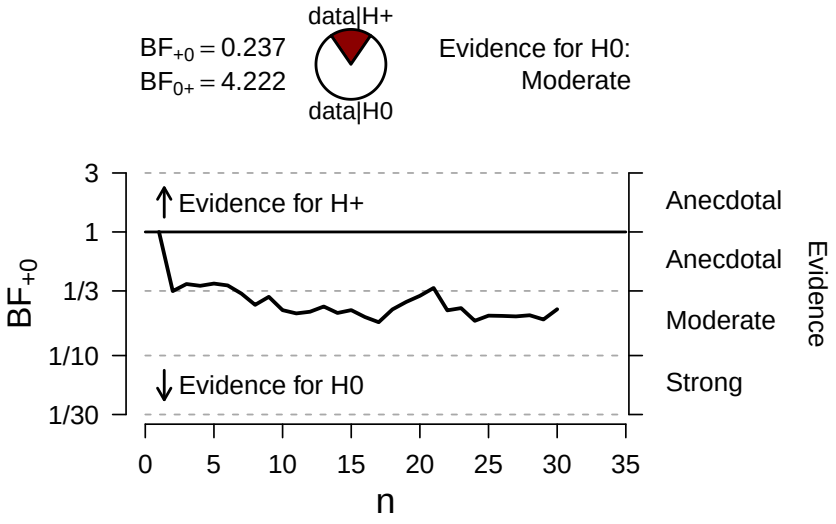


Figure 2.3: Sequential analyses for Experiment 1 for the speech stimuli in the generalization conditions with LTs longer than 15 seconds allowed.

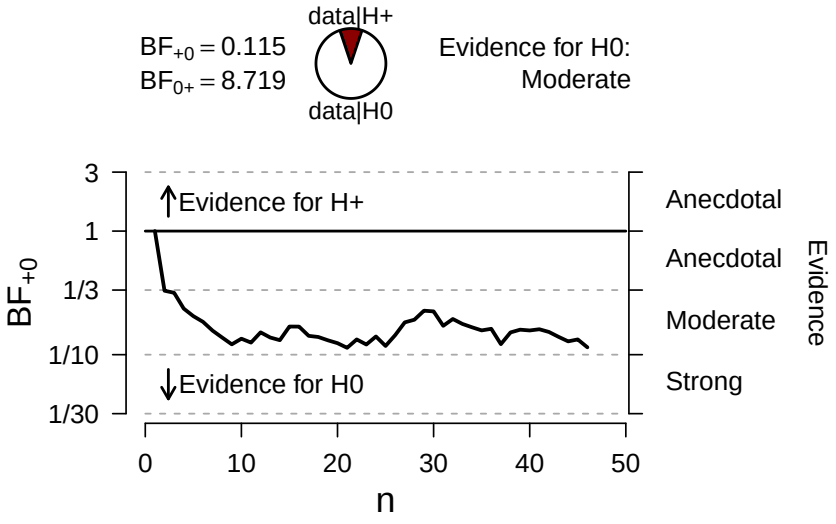


Figure 2.4: Sequential analyses for Experiment 1 for the song stimuli in the generalization conditions with LTs longer than 15 seconds allowed.

more likely than the alternative hypothesis as all BFs are greater than 1. Also note that the highest BFs can be found in the cells with lowest p-values. This may seem counter-intuitive, but is the case because the small effects that are found are in the opposite direction. Infants look longer when consistent patterns are presented than when inconsistent patterns are presented, providing more evidence against a preference for inconsistent sounds.

Table 2.3: Bayes Factors that indicate how much more likely H0 (that infants do not have a preference) is than H1 (that infants have a preference for inconsistent sounds), presented for three conditions (Combination, Generalization, Place), two types of stimuli (Speech, Song), and two types of preprocessing options (with and without looking times longer than 15 seconds).

	Combination	Generalization	Place
Speech, all LT	9.23	4.22	5.47
Song, all LT	4.70	8.72	18.38
Speech, LT < 15	9.56	4.22	5.78
Song, LT < 15	4.31	10.26	15.08

2.3.3 Discussion

The analysis did not show evidence of differential looking to consistent or inconsistent test patterns overall except in the specific case of Place test items in the Song condition if LTs over 15 seconds were allowed. If those long looking times were removed the effect was no longer found, indicating that it was not a strong effect to begin with. No other main effects were found.

There were, however, interaction effects between the familiarization pattern and the consistency of the test pattern. Infants preferred the XXY pattern, independent of the pattern with which they were familiarized.

Counter to our initial hypotheses, we did not find evidence for better performance when infants were exposed to speech, nor did we find any effect of development. We recognized that the testing conditions presented in this experiment may have placed an unrealistically high cognitive load on the infants and that this may have obscured any potential evidence of learning. Including test items with various degrees of difference from the familiarization items, and in unequal numbers, may have been ecologically valid, as infants hear a variety of familiar and novel items intermixed on a daily basis, but the amount of variety may have confused them, not allowing them to show differential attention on the basis of pattern alone. We thus hypothesized that if the testing sets were simpler and included only generalization test items, as in Marcus et al. (1999), infants might be more prone to show differential looking between test items consistent or inconsistent with the familiarization pattern.

2.4 Experiment 2

2.4.1 Methods

Stimuli

The stimuli in Experiment 2 differed from those used in Experiment 1 only with respect to the test items used. In this experiment, only the generalization items, GGH, GHG, HHG, and HGH were used in test (see Table 2.2 and Appendix B2). Because we found no difference with respect to the use of speech or song in Experiment 1, during this experiment, only speech stimuli were used.

Participants

We tested 25 six-month-olds and 22 nine-month-olds. Recruitment and compensation was identical to Experiment 1.

Infants were removed from analysis for the following reasons: technical errors (n=5) and excessive fussiness (n=1). Ten of the infants in this experiment were exposed to another language in addition to Dutch, with all infants having one Dutch parent and being exposed to Dutch both in- and outside the home.

Procedure

The procedure was identical to Experiment 1, with the exception of the type and number of test trials. Four generalization test items were used, two that were consistent with the training pattern, and two that were inconsistent with the training pattern. These four test trials were repeated three times, for a total of 12 test trials (cf. Marcus et al., 1999).

2.4.2 Results

Data preprocessing

As in Experiment 1, LTs that were shorter than the duration of one triad (1.5 seconds) were excluded (4% of the data). Again, there were LTs that exceed the 15-second presentation time of the stimuli (6% of the data) and we report the analyses both with and without these LTs.

Frequentist analysis

The multiverse analysis (Figure 2.5) did not find any differences between including or omitting LTs longer than 15 seconds. In both cases, there were no significant effects of consistency, but there was an interaction effect between training and consistency, similar to the interaction effect found in Experiment 1. In order to interpret this interaction effect, the mean LTs for the consistent and inconsistent test stimuli and familiarization stimuli are visualized in Figure

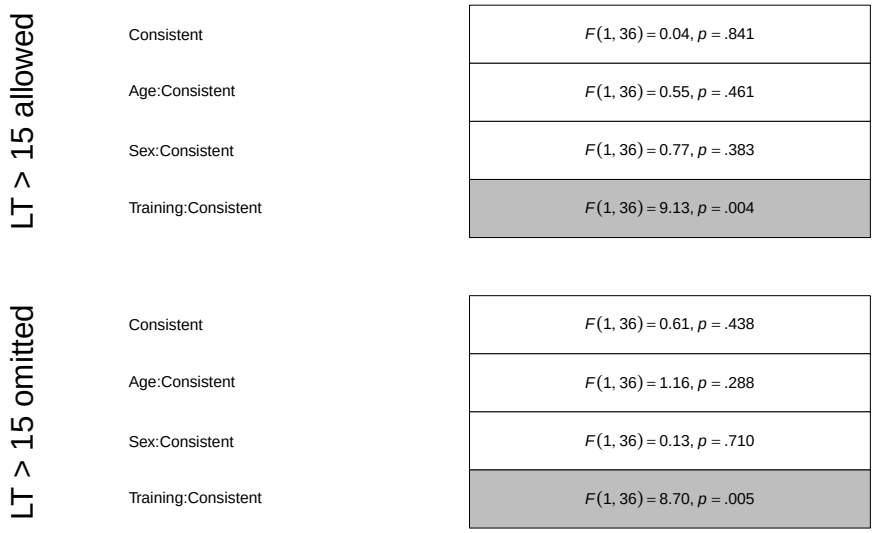


Figure 2.5: Multiverse of p-values for Experiment 2.

2.6 with error bars of ± 1 standard error. As in Experiment 1, infants seemed to have a preference for patterns including adjacent repetitions (XXY).

Bayesian analysis

To help our understanding of the null effect we also performed a sequential Bayesian analysis, using BFs to quantify how much more likely the null hypothesis is than the alternative hypothesis. Figures 2.7 and 2.8 shows how evidence accumulates in favor of the null hypothesis for both the data with and without LTs longer than 15 seconds included.

2.4.3 Discussion

Infants in Experiment 2, like those in Experiment 1, did not show differential looking between consistent and inconsistent test items overall. As in Experiment 1, we found an interaction between the familiarization pattern and consistency of the testing pattern. This result indicates that in Experiment 2, infants also preferred to look more to XXY test items during the test phase,

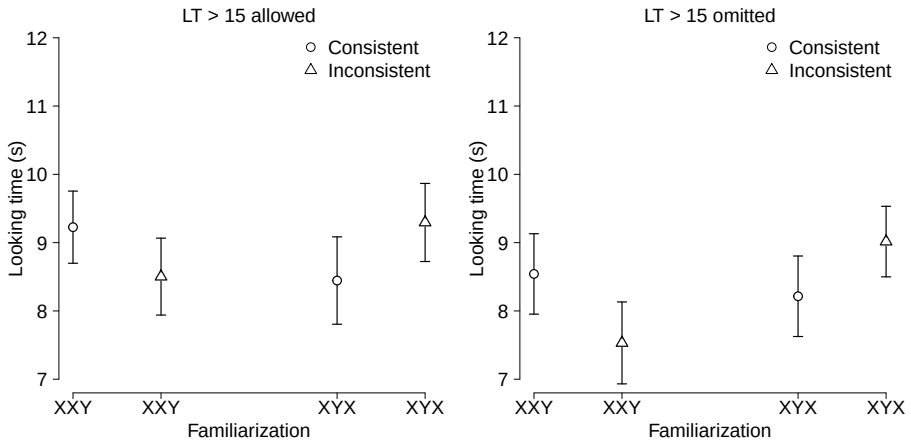


Figure 2.6: Interactions of familiarization and test grammars in Experiment 2.

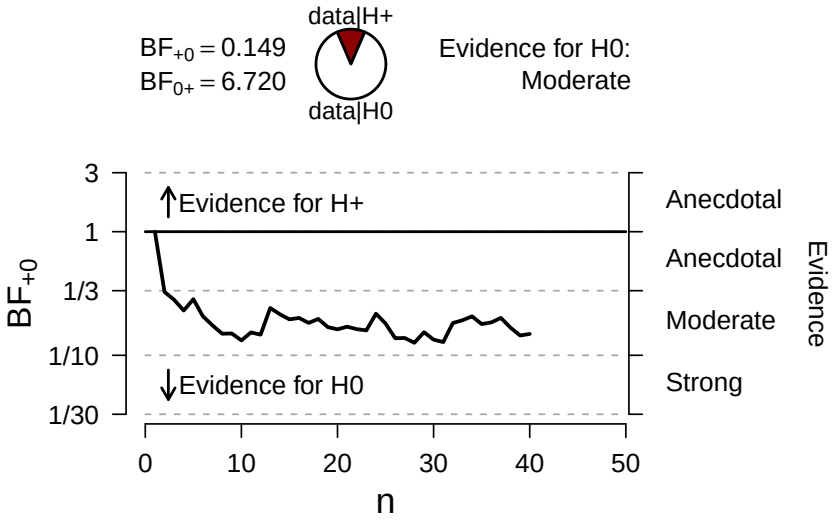


Figure 2.7: Sequential analyses for Experiment 2 with LTs longer than 15 seconds allowed.

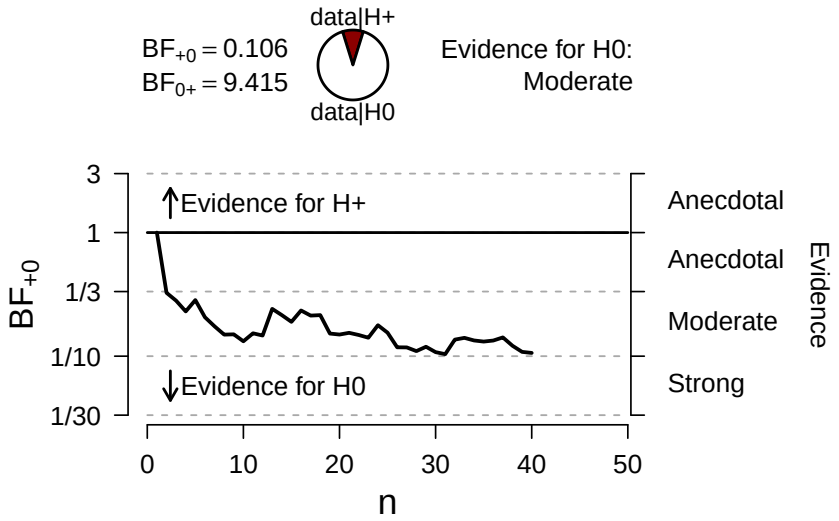


Figure 2.8: Sequential analyses for Experiment 2 with LTs longer than 15 seconds not allowed.

independent of which pattern they had been exposed to. While this is an interesting result consistent with the idea that infants may have a bias for adjacent repetitions, it does not replicate the results of Marcus et al. (1999) and does not show clear evidence of the ability to learn rules.

After failing to show rule learning in two experiments using the same familiarization phase and procedure, we identified several points on which our experiment differed from the original Marcus et al. (1999) study that may have negatively impacted the ability of infants to learn or to show their learning. First, our carefully controlled stimuli may have introduced a high level of variability which did not allow the infants to learn the rule quickly enough. This may have been due to the fact that our stimuli were phonologically complex as compared to those used in Marcus et al. The presence of more heterorganic syllable patterns in our stimuli might have attracted attention to the syllable structure rather than to the pattern.

A second reason for infants' failure to generalize may have been the fact that our patterns were constructed from uncategorized elements – any syllable could appear in any position, except for specifically the "F" syllable which changed position in Experiment 1, from appearing exclusively in the Y position during familiarization to the X position during test. We know from literature that categorization allows learners to attend to more than just the individual syllables and to form a rule. Braine (1987; summarized in Gerken, Wilson, &

Lewis, 2005) showed that the first step in solving an AGL task was to learn that there were categories in the grammar. Gómez and Gerken (2000) also argued that being able to make abstract categories is crucial for language learning and productivity.

In addition to the differences with respect to the stimuli, we used a different experimental paradigm. While VFP and HPP are based on the same principle of looking-while-listening, there may be differences with respect to how active the infant has to be in each task, the HPP being slightly more active (details below).

With these considerations in mind we ran a third experiment to try to replicate Marcus et al. (1999) as closely as possible, using similar stimuli and a similar procedure.

2.5 Experiment 3

2.5.1 Methods

Stimuli

In this experiment, the stimuli were similar to those used in Marcus et al.'s (1999) Experiments 2 and 3, with two small exceptions. As our infant participants were Dutch native speakers, a Dutch (rather than an American English) synthesizer was used to create the syllables (synthesized in Praat; Boersma & Weenink, 2007). In addition, the voice affricate /dʒ/ present in syllables *ji* and *je* does not exist in Dutch. Because there are no voice affricates in Dutch, these syllables were replaced with with voiced fricative /v/ to produce *vi* and *ve*. The triads used are shown in Table 3.

Participants

We tested 30 monolingual Dutch seven-month-old infants. Recruitment and compensation was identical to Experiments 1 and 2.

Infants were removed from analysis for the following reasons: technical errors (n=5), excessive fussiness (n=6), being born more than 3 weeks preterm (n=1), and parental decision to stop the experiment (n=2).

Procedure

Infants were tested in the HPP as in Marcus et al. (1999). As in Experiments 1 and 2, infants were placed on their caregiver's lap in a sound attenuated booth. The caregiver wore headphones playing non-rhythmical music from a female vocal artists to mask the auditory stimuli of the experiment. At a distance of 110 cm on each side of the room was one circular, red LED light (the side lights). In the center, directly in front of the infant at a distance of 110 cm was a circular, green LED light (the attention grabber). Directly below the attention

Table 2.4: Familiarization and test items following the XYY and XYX pattern respectively.

XYY	XYX
Familiarization	
dedidi	dedide
deveve	devede
delili	delide
dewewe	dewede
vididi	vidivi
viveve	vivevi
vilili	vilivi
viwewe	viwevi
ledidi	ledile
leveve	levele
lelili	lelile
lewiwi	lewile
wididi	widiwi
wiveve	wivewi
wilili	wiliwi
wiwewe	wiwewi
Test	
bapopo	bapoba
kogaga	kogako

grabber was a wide-angle video camera (Go Pro Hero 3), which allowed the experimenter to view and react to the infant's behavior and which recorded the experiment for off-line coding. Directly above the side lights were speakers, hidden below white sheets of fabric to minimize distraction away from the lights.

The experiment began with an attention grabber. As in the previous two experiments, when the infant focused on the attention grabber for two seconds, the familiarization phase began. During this phase, the familiarization stream, consisting of 16 triads repeated three times in different randomizations, played uninterrupted for approximately two minutes (1 min 46 sec). In addition, the side lights were illuminated when the infants looked at them and were extinguished when the infant looked away, in order to familiarize the infants with the fact that the lights were contingent on their attention. This was done independently of the auditory stream.

When the familiarization phase ended, the attention grabber played until the infants again looked towards it for a total of two seconds. Then the test phase began with one of the side lights flashing. The side of the first light was counterbalanced between infants. As soon as the infants turned their heads

towards the flashing side light, the auditory test stimuli started to play. As in Experiments 1 and 2, the test stimuli would continue playing as long as the infants looked, for a maximum of 15 seconds. When the infants looked away for less than two seconds, the auditory stimuli would briefly pause, but would resume if the infants turned back within that time window. If the infants looked away for more than two seconds, the auditory stimuli and the blinking light would stop, the trial would end, and the attention grabber would begin again. This repeated until the infants finished the test phase, or until the caretaker or experimenter determined that an infant was too fussy (not orienting to attention grabber for an extended period of time) or distressed (crying or screaming while ignoring the lights). As in Experiment 2, the test phase consisted of four generalization test items, two corresponding to a pattern consistent with the familiarization and two corresponding to a pattern inconsistent with the familiarization. These four test items were repeated for three blocks, and randomized per block, totaling 12 test trials.

2.5.2 Results

Data preprocessing

As this was a replication, we used the same procedure as Marcus et al. (1999), LTs shorter than the duration of the stimuli were omitted (3% of the data) and LTs longer than 15 seconds were kept (16% of the data).

Frequentist analysis

A repeated measures ANOVA with consistency, training and their interaction on the mean looking times did not yield a significant main effect for consistency [$F(1,14) = 1.74$, $p = .208$)]. There was a near-significant interaction effect between training and condition [$F(1,14) = 4.05$, $p = .064$, see Figure 2.9]. Infants who were trained with XYX had a preference for inconsistent trials, whereas infants trained with XYY had a preference for consistent trials. This implies an overall preference for XYY over XYX, similar to the results of Experiments 1 and 2, where there was a preference for XXY over XYX.

Bayesian analysis

As in the previous two experiments, we performed a sequential Bayesian paired samples t-test to quantify how much more likely the null hypothesis is than the alternative hypothesis that infants do look longer when inconsistent sound are played. Figure 2.10 shows how evidence accumulates in favor of the null hypothesis.

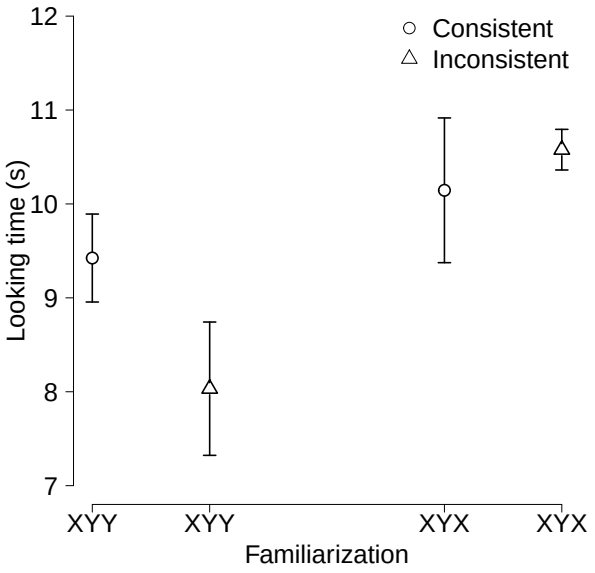


Figure 2.9: Interactions of familiarization and test grammars in Experiment 3.

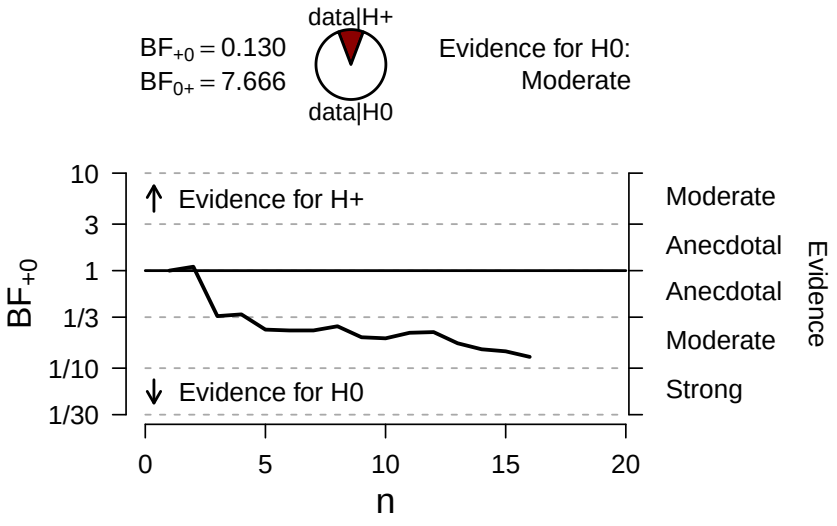


Figure 2.10: Sequential analyses for Experiment 3.

2.5.3 Discussion

Counter our expectations, infants in this experiment did not perform differently from those in the previous two experiments: they did not show differential looking based on consistency of test items with the familiarization grammar, indicating that they did not generalize the rule to novel input. They did, however, show longer looking to the test pattern containing adjacent repetition, the XYY pattern in this case. Despite using the similar syllables and triads, the same procedure, and the same age group as in the original Marcus et al. (1999) experiment, we could not replicate the results with Dutch infants.

2.6 Combination of the 3 experiments

Taken together, these three experiments seem to imply that infants do not have a preference for inconsistent over consistent patterns, but that they do have a preference for repeating patterns over alternating patterns. To illustrate these two points we performed a combined Bayesian analysis with the data of the three experiments combined. Using a Bayesian approach is especially well suited, as the outcomes can be framed as competing hypotheses, allowing us to combine data from multiple experiments.

2.6.1 Results

Data preprocessing

Although there are many possible options to preprocess the data, the multiverse analyses (Steegen et al., 2016) of Experiments 1 and 2 indicated most choices did not matter. Therefore, we kept the combined analyses close to the original Marcus et al. (1999) paper and use the same choices made in Experiment 3. That is, LTs longer than 15 seconds were kept while LTs shorter than the duration of one triad were omitted.

Bayesian analysis - Overall

Figure 2.11 shows how evidence accumulates in favor of the null hypothesis. Together the three experiments provide strong evidence in favor of the null hypothesis that infants have no preference for either consistent or inconsistent sounds. The data of 132 infants shows that the null hypothesis is approximately 14 times more likely than the alternative hypothesis that infants have a preference for inconsistent patterns.

Bayesian analysis - Repetition

For all of the three experiments we calculated the mean LTs during the test phase for the repeating patterns (XXY and XYY) and the non-repeating pattern (XYX). These LTs were then used to perform a sequential Bayesian paired

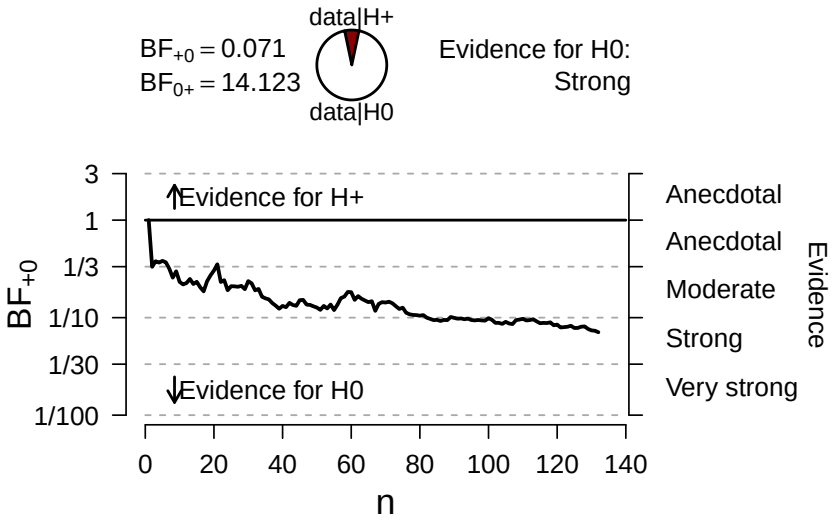


Figure 2.11: Sequential analyses for all three experiments showing evidence for the null hypothesis.

samples t-test. Figure 2.12 shows how evidence accumulates in favor of the repeating patterns hypothesis. Together the three experiments provide extremely strong evidence in favor of a preference for repeating over non-repeating patterns. The data of 134 infants shows that a preference for repeating patterns is 1177 times more likely than no preference or a preference for non repeating patterns.

2.7 General Discussion

Across three experiments with over 100 infants, we attempted to extend on, and eventually simply to replicate, the work of Marcus et al. (1999). The original study had shown that seven-month-old infants are able to learn simple XYX, XYY, and XXY rules on an abstract level, generalizing them to novel stimuli. In the original study, infants showed significantly more interest in test patterns that were inconsistent with the familiarization pattern than those that were consistent with it, independent of which familiarization pattern was used. In our experiments, Bayesian statistical analyses showed moderate evidence in favor of the null hypothesis in each experiment, indicating that infants did not discriminate the familiarized pattern from the novel pattern based on consistency. In Experiments 1 and 2, infants trained on XXY looked longer to the consistent XXY, while infants trained on XYX looked longer to the inconsis-

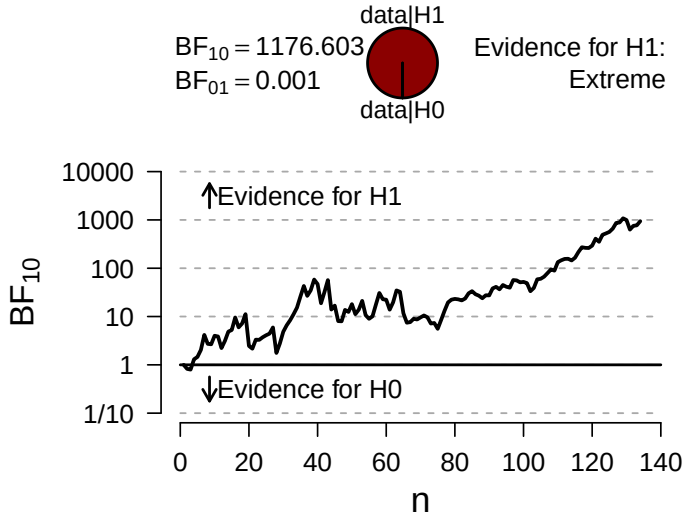


Figure 2.12: Sequential analyses for all three experiments showing evidence for the alternative hypothesis that infants look differently based on the presence of repetition.

tent XXY. Similarly, in Experiment 3, infants trained on XYY preferred the consistent XYY test pattern, while infants trained with XYX preferred the inconsistent XYY test pattern. We interpret the difference in looking towards the consistent and inconsistent patterns based on the familiarization pattern to indicate a general preference for patterns containing adjacent repetition.

Indeed, a Bayesian analysis of all the data confirmed that there is strong evidence for the null hypothesis that infants have no preference based on consistency of the test pattern with the familiarization pattern, and extremely strong evidence for the alternative hypothesis that infants are more likely to prefer adjacent repetition than to have no preference. The current set of studies is a good illustration of the practical benefits of a Bayesian analysis approach. First, we were able to quantify the plausibility of the null hypothesis that infants have no preference for either consistent or inconsistent sounds. Furthermore, the presented cumulative analyses show how incrementing the amount of data strengthens conclusions. The latter feature of Bayesian analysis introduces a new way of conducting infant experiments, which would entail testing until either the null-hypothesis or the alternative hypothesis gathers enough evidence. This would constitute a practical solution to the expensive and complicated practice of infant research.

2.7.1 Repetition biases

In our experiments, we found evidence for a repetition bias in infants. If infants have a general preference for repetition patterns, we might expect to find that they show this preference already during familiarization – infants trained on XXY might show longer looking times during familiarization than infants trained on XYX. Yet no such preference was found. One explanation may be that the preference is not found during familiarization simply because of the nature of the familiarization paradigm: the auditory stimulus continues independently of infant's interest in all three experiments such that even if they were not interested in the stimuli, infants had no incentive to modulate their behavior accordingly and show their preference or lack thereof. On the other hand, it might also be indicative that the preference simply does not manifest itself except in relation to the familiarization pattern. When presented with a reference point first, infants later prefer the test items that are similar or different based on the presence of repetition[6]. Such an explanation *would* require learning at some level, at least learning of the presence or absence of repetition. In either case, the conclusion we can draw from this work is not that speech is special but that repetition is special.

This type of bias for repetitions had been shown in previous literature but the preference has manifested itself in a different way. Most notably, Gervain et al. (2008) found that newborn infants' brains already react differently to the presence of repetitions in sound sequences than to the absence of repetitions. In addition, there is evidence that the presence of repetitions is favorable for rule learning. In the visual domain, Johnson et al. (2009) found that 11-month-old infants are able to learn XYY and XXY rules, but not XYX rules; eight-month-old infants were only able to learn the XYY rule. These findings show that there is a hierarchy of difficulty with respect to even these very simple patterns. Our results line up with these empirical findings on infants, with cross-species findings showing a bias for repetitions (Chen et al., 2015; van Heijningen et al., 2013), and with the theoretical arguments of Endress, Mehler, and colleagues who identify repetitions, particularly at edges, as one of the main biases that aids in learning structures and patterns across domains (Endress et al., 2009; Endress, 2013). While we could not conclude that algebraic rule learning was occurring in our experiments, the results provide evidence for a preferential status of repetitions in perception in the first months of life. From our results, we conclude that the ability to learn and generalize simple Marcus rules may be far more fragile than previously thought, even in the speech domain (whether natural or synthetic).

[6] A small pilot experiment in our lab conducted without a familiarization phase did not reveal a preference for one type of pattern over another either, adding support for this account.

2.7.2 Future work

The results raise several possible paths for future work. First, a possible reason for discrepancy between Experiment 3 and the original Marcus et al. (1999) study could be the difference in native language of the participants. In Experiments 1 and 2, we used natural Dutch syllables that did not produce real words. In Experiment 3, except for the replacement of one consonant, we used the stimuli from Marcus et al. (1999) produced by a Dutch synthesizer. The frequency distribution of the particular syllables used in the original study may have been different in English than in Dutch. On a related note, the similarity between phonemes and between syllables may be different in English than in Dutch. Simple perceptual cues and frequencies may play a role in the ability to notice different patterns, may modulate attention, and thus may influence learning ability during the familiarization phase. How similar or different individual items in a to-be-learned triad should be is worth exploring further in the quest to understand what conditions do and do not allow learning to occur. However, if learning outcomes indeed turn out to be highly dependent on these cues, it would show again that this type of rule learning is fragile, and one could question whether this type of algebraic rule learning is possible outside the lab at all.

This begs the question of whether these types of rule learning tasks are actually successfully tapping into learning abilities that we know infants have. The role of artificial grammar learning is to shed light on the cognitive building blocks that are at work in allowing for rule learning to occur in natural language. Yet we must ask whether these types of patterns are successful tools in researching these abilities. While the patterns XXY and XYY do occur as reduplication at different levels of language (word-internally, and within and across sentences), the pattern XYX in is more unrealistic as a representation of anything occurring in natural language at the word or sentence level[7,8]. In attempting to interpret the theoretical consequence of infants' failure to learn such unnatural patterns, we are left with the question of why infants cannot do something simple when we know they can do far more complicated computations and abstractions. Future work might move away from the use of these Marcus patterns and focus on more complex yet still simple structures from which clearer parallels with natural language can be drawn.

Acknowledgements

We would like to thank all who helped in setting up appointments and collecting data: Marijn van 't Veer, Renske Jacobs, Rabia Mahboeb, Astrid Gilein,

[7]Examples of exceptions in English would be phrases such as *Dog eat dog*, *Dust to dust*, *Brother against brother*, etc. Idiomatic expressions and phrases aside, this remains a rarely found phenomenon.

[8]See also chapter 4 for an analysis of differences between English and Dutch at the level of the three-syllable patterns used here.

Mirjam Bahlmann, Isabella Jordanoska, Gijsbert Westland. We thank Reinhard Schillberg, Vincent Buurman, and Saskia Lensik for analysis advise along the way. We are very grateful to Carel ten Cate for comments on this paper, and to Michelle Spierings, Carel ten Cate, Raquel Garrido Alhama, and Jelle Zuidema for discussion and feedback throughout this project. The research reported here was supported by NWO Vrije Competitie grant 360.70.452 to Clara C. Levelt.

2.8 Appendices

2.8.1 Appendix A

Table A1: Each letter within the triads presented in Table 2 corresponded to a different syllable for each participant according to the following scheme.

	Participant number							
Syllable label	1	2	3	4	5	6	7	8
A	do	fi	ka	le	mo	nu	pu	sa
B	fi	ka	le	mo	nu	pu	sa	do
C	ka	le	mo	nu	pu	sa	do	fi
D	le	mo	nu	pu	sa	do	fi	ka
E	mo	nu	pu	sa	do	fi	ka	le
F	nu	pu	sa	do	fi	ka	le	mo
G	pu	sa	do	fi	ka	le	mo	nu
H	sa	do	fi	ka	le	mo	nu	pu

2.8.2 Appendix B

Table B1: Familiarization triads for Experiments 1 and 2, per grammar.

Familiarization Grammar XXY								
	1	2	3	4	5	6	7	8
AAD	do do le	fi fi mo	ka ka nu	le le pu	mo mo sa	nu nu do	pu pu fi	sa sa ka
AAE	do do mo	fi fi nu	ka ka pu	le le sa	mo mo do	nu nu fi	pu pu ka	sa sa le
AAF	do do nu	fi fi pu	ka ka sa	le le do	mo mo fi	nu nu ka	pu pu le	sa sa mo
BBA	fi fi do	ka ka fi	le le ka	mo mo le	nu nu mo	pu pu nu	sa sa pu	do do sa
BBC	fi fi ka	ka ka le	le le mo	mo mo nu	nu nu pu	pu pu sa	sa sa do	do do fi
BBF	fi fi nu	ka ka pu	le le sa	mo mo do	nu nu fi	pu pu ka	sa sa le	do do mo
CCA	ka ka do	le le fi	mo mo ka	nu nu le	pu pu mo	sa sa nu	do do pu	fi fi sa
CCD	ka ka le	le le mo	mo mo nu	nu nu pu	pu pu sa	sa sa do	do do fi	fi fi ka
CCF	ka ka nu	le le pu	mo mo sa	nu nu do	pu pu fi	sa sa ka	do do le	fi fi mo
DDB	le le fi	mo mo ka	nu nu le	pu pu mo	sa sa nu	do do pu	fi fi sa	ka ka do
DDE	le le mo	mo mo nu	nu nu pu	pu pu sa	sa sa do	do do fi	fi fi ka	ka ka le
DDF	le le nu	mo mo pu	nu nu sa	pu pu do	sa sa fi	do do ka	fi fi le	ka ka mo
EEB	mo mo fi	nu nu ka	pu pu le	sa sa mo	do do nu	fi fi pu	ka ka sa	le le do
EEC	mo mo ka	nu nu le	pu pu mo	sa sa nu	do do pu	fi fi sa	ka ka do	le le fi
EEF	mo mo nu	nu nu pu	pu pu sa	sa sa do	do do fi	fi fi ka	ka ka le	le le mo

Familiarization Grammar XYX								
	1	2	3	4	5	6	7	8
ADA	do le do	fi mo fi	ka nu ka	le pu le	mo sa mo	nu do nu	pu fi pu	sa ka sa
AEA	do mo do	fi nu fi	ka pu ka	le sa le	mo do mo	nu fi nu	pu ka pu	sa le sa
AFA	do nu do	fi pu fi	ka sa ka	le do le	mo fi mo	nu ka nu	pu le pu	sa mo sa
BAB	fi do fi	ka fi ka	le ka le	mo le mo	nu mo nu	pu nu pu	sa pu sa	do sa do
BCB	fi ka fi	ka le ka	le mo le	mo nu mo	nu pu nu	pu sa pu	sa do sa	do fi do
BFB	fi nu fi	ka pu ka	le sa le	mo do mo	nu fi nu	pu ka pu	sa le sa	do mo do
CAC	ka do ka	le fi le	mo ka mo	nu le nu	pu mo pu	sa nu sa	do pu do	fi sa fi
CDC	ka le ka	le mo le	mo nu mo	nu pu nu	pu sa pu	sa do sa	do fi do	fi ka fi
CFC	ka nu ka	le pu le	mo sa mo	nu do nu	pu fi pu	sa ka sa	do le do	fi mo fi
DBD	le fi le	mo ka mo	nu le nu	pu mo pu	sa nu sa	do pu do	fi sa fi	ka do ka
DED	le mo le	mo nu mo	nu pu nu	pu sa pu	sa do sa	do fi do	fi ka fi	ka le ka
DFD	le nu le	mo pu mo	nu sa nu	pu do pu	sa fi sa	do ka do	fi le fi	ka mo ka
EBE	mo fi mo	nu ka nu	pu le pu	sa mo sa	do nu do	fi pu fi	ka sa ka	le do le
ECE	mo ka mo	nu le nu	pu mo pu	sa nu sa	do pu do	fi sa fi	ka do ka	le fi le
EFE	mo nu mo	nu pu nu	pu sa pu	sa do sa	do fi do	fi ka fi	ka le ka	le mo le

Table B2: Test triads. For Experiment 1, all items were used, while in Experiment 2 only triads labeled "Generalization" were used.

		Test Phase							
		1	2	3	4	5	6	7	8
Combination	AAC	do do ka	fi fi le	ka ka mo	le le nu	mo mo pu	nu nu sa	pu pu do	sa sa fi
	ACA	do ka do	fi le fi	ka mo ka	le nu le	mo pu mo	nu sa nu	pu do pu	sa fi sa
	BBD	fi fi le	ka ka mo	le le nu	mo mo pu	nu nu sa	pu pu do	sa sa fi	do do ka
	BDB	fi le fi	ka mo ka	le nu le	mo pu mo	nu sa nu	pu do pu	sa fi sa	do ka do
	CCE	ka ka mo	le le nu	mo mo pu	nu nu sa	pu pu do	sa sa fi	do do ka	fi fi le
	CEC	ka mo ka	le nu le	mo pu mo	nu sa nu	pu do pu	sa fi sa	do ka do	fi le fi
	DDC	le le ka	mo mo le	nu nu mo	pu pu nu	sa sa pu	do do sa	fi fi do	ka ka fi
	DCD	le ka le	mo le mo	nu mo nu	pu nu pu	sa pu sa	do sa do	fi do fi	ka fi ka
	EEA	mo mo do	nu nu fi	pu pu ka	sa sa le	do do mo	fi fi nu	ka ka pu	le le sa
	EAE	mo do mo	nu fi nu	pu ka pu	sa le sa	do mo do	fi nu fi	ka pu ka	le sa le
Position	FFA	nu nu do	pu pu fi	sa sa ka	do do le	fi fi mo	ka ka nu	le le pu	mo mo sa
	FAF	nu do nu	pu fi pu	sa ka sa	do le do	fi mo fi	ka nu ka	le pu le	mo sa mo
	FFE	nu nu mo	pu pu nu	sa sa pu	do do sa	fi fi do	ka ka fi	le le ka	mo mo le
	FEF	nu mo nu	pu nu pu	sa pu sa	do sa do	fi do fi	ka fi ka	le ka le	mo le mo
Generalization	GGH	pu pu sa	sa sa do	do do fi	fi fi ka	ka ka le	le le mo	mo mo nu	nu nu pu
	GHG	pu sa pu	sa do sa	do fi do	fi ka fi	ka le ka	le mo le	mo nu mo	nu pu nu
	HHG	sa sa pu	do do sa	fi fi do	ka ka fi	le le ka	mo mo le	nu nu mo	pu pu nu
	HGH	sa pu sa	do sa do	fi do fi	ka fi ka	le ka le	mo le mo	nu mo nu	pu nu pu

CHAPTER 3

Rule learning in infants in the visual domain

This chapter is based on a paper in preparation for submission as Geambaşu, A., van Leeuwen, C., & Levelt, C.C. (in preparation). Lexical knowledge and rule learning in infants.

The chapter addresses the domain-specificity question in infant rule learning by investigating rule learning abilities in the visual domain. It incorporates infants' lexical knowledge as a potential facilitatory factor in rule learning.

3.1 Abstract

To what extent abstract, algebraic rule learning of the type proposed by Marcus et al. (1999) is domain-specific or not is relevant for understanding the process that is at the foundation of our ability to learn the grammatical rules of language. Here we present an experiment conducted with 12- to 14-month-olds in which we familiarized the infants with XYY and XYX patterns in the same manner as in Saffran et al. (2007), in three different conditions. Infants were presented with sequences of three pictures of inanimate objects for which they either knew the label (Known condition), of familiar items for which they did not know the label (Unknown condition), or of nonsense items (Nonsense condition). The pictures were presented in either an XYY or an XYX pattern during familiarization. During the test, novel Known, Unknown or Nonsense items were presented, to ensure generalization of the rule beyond the familiarization input, in sequences that were either consistent or inconsistent with the training sequence. Our results show a failure to replicate the effects of

Saffran et al. (2007). Infants 12- to 14-months did not show evidence of differential looking with respect to the test grammar's consistency with the training grammar overall. Only in one very specific subset of participants did we find evidence of generalization: boys in the Known condition trained with the XYY grammar showed a consistency preference. All infants in the Known condition also looked significantly more overall throughout the experiment than infants in the Unknown or Nonsense conditions. While our experiments show a failure to replicate the findings of Saffran et al. (2007), they help identify points of focus for future work relating to differences between sexes and the role of lexicon, minimal differences between stimuli, and the salience of repetitions in rule learning.

3.2 Introduction

The ability to learn patterns and extract "algebraic" rules (Marcus et al., 1999) has been explored in the speech domain as a feature underpinning our ability to learn the complex syntactic rules of language. It has also been explored outside the speech domain, as a potential domain-general cognitive ability. While some authors argue that speech is special with respect to our ability to learn and generalize abstract rules, others have proposed that a single cognitive mechanism allows us to learn rules in multiple domains, including language. Yet infants' rule learning experiments with non-linguistic stimuli, such as tones or visual stimuli, have shown contradictory results.

Marcus et al. (2007) are among the researchers who argue that rule learning is facilitated when speech stimuli are used. These authors found that infants could not learn rules if the familiarization stimuli consisted of non-speech sounds. Another study from some of the same authors (Johnson et al., 2009) has shown that rule learning in the visual domain is possible but restricted, depending on properties of the patterns themselves and on the infant's cognitive development. The authors found that eight-month-old infants learned the rule only if they were familiarized with an XYY pattern and were exposed to XYY and XYX in the test phase. In contrast, 11-month-old infants were able to discriminate both grammars containing immediate repetitions from each other (XYY vs. XXY and XXY vs. XYY) as well as the early immediate repetition from a non-adjacent repetition (XXY vs. XYX). Both eight month olds and 11 month olds trained on XYX were unable to discriminate this grammar from an XYY grammar. These results show that rule learning outside of the speech domain may be less robust, and suggest that, as Marcus et al. (2007) claimed, speech may be special with respect to rule learning.

Saffran et al. (2007), however, showed that 7.5-month-old infants were able to learn rules from visual stimuli when the stimuli consisted of pictures of cats and dogs (replicated by Ferguson & Waxman, 2015). Saffran and colleagues argued that rule learning was thus not speech-specific or language-specific, and that the success of the seven-month-old infants in their study had to do with

the familiarity of the pictured objects, which facilitated their categorizability. Categorization has long been argued to be beneficial for artificial grammar learning (Brain, 1987; Gerken et al., 2005; Gómez & Gerken, 2000). In contrast, according to Saffran, items used in previous studies with shapes had shown more limited results (Marcus et al., 2004[1] as cited in Saffran, 2007) because infants are not able to easily identify and categorize shapes. Dogs and cats, on the other hand, are both familiar and interesting to them and can be categorized – i.e., a pattern of, for example, Dog X, Dog Y, Dog X can be discerned. This allowed them to learn and generalize the XYX or XYY pattern equally well to novel instances of dog or cat breeds.

This familiarity effect may also help to explain asymmetric results in other visual rule learning experiments. For example, when hearing 7.5-month-old infants were exposed to XYY and XXY rules composed of American Sign Language signs, they were able to generalize the XYY rule, but not the XXY rule, to novel input (Rabagliati et al., 2012). These hearing infants had never been exposed to these structured signs before, and were therefore not familiar with them (and probably did not recognize them as being linguistic units either). In this case, infants seemed to mediate the difficulty of learning the task with these unknown stimuli by attending to a final repetition pattern, which is a highly salient perceptual primitive that aids in rule learning (Mehler et al., 2008; Endress et al., 2009). Thus, familiarity with the stimuli may be an important factor in aiding rule learning, and in the case of unfamiliar stimuli, other biases step in.

However, an alternative explanation to the familiarity effect may be that language did still play a role in infants' success in the Saffran et al. (2007) experiment. The infants may have been able to learn the abstract rules because they had a lexical representations of the visual stimuli that were used in the experiment. Infants at 7.5 months are likely to have lexical representations for *dog* and *cat* - these names are on the Dutch Communicative Development Inventory (NCDI) level 1 questionnaire (and also see Bergelson & Swingley 2012). This lexical knowledge may have aided them in committing the structure to memory and, in turn, in shifting their attention beyond the item level to the rule level. In contrast, infants at this age will not have labels referring to abstract shapes like triangles and squares - these items are *not* on the NCDI level 1 questionnaire. Success in rule learning with pictures of dogs and failure in rules learning with pictures of abstract shapes might therefore be due to the presence or absence of a lexical representation for the stimuli, rather than to the familiarity of the items per se.

The work presented here thus rests on the hypothesis that it is infants' lexical knowledge that allowed them to learn abstractions of the XYX and XYY rules presented in the work of Saffran and colleagues. If this is the case, we expect infants to learn better (or perhaps only) when they are presented with sequences of pictures they can label passively (i.e., comprehend), and

[1]Unpublished conference presentation.

that they have more difficulty (or fail completely) when they are presented with sequences of pictures of items that they cannot yet label.

To test this hypothesis we created XYX and XYY grammars in three conditions. In the first condition (Known) stimuli consisted of familiar items that 12-14 month olds could (very probably) label. In the second condition (Unknown), infants were presented with familiar items for which they probably did not have a lexical representation yet. In the final condition (Nonsense), infants were presented with unfamiliar, nonsense items they could not have seen before, that do not have lexical labels. With these three conditions, we attempted to better understand the "familiarity" effect, that is, whether the facilitatory effect results from bare item familiarity, or from the presence of a lexical representation for the item, i.e. linguistic familiarity.

Our results show significantly longer looking times overall, yet limited evidence of rule learning in the Known condition, with different performances between boys and girls and between familiarization grammars. There was no evidence of rule learning in the Unknown and Nonsense conditions; data in these conditions varied due to various individual differences.

3.3 Method

3.3.1 Participants

We tested 89 infants between the ages of 12 and 14 months. Infants were recruited in and around the city of Leiden, The Netherlands, via the local municipality and came mostly from monolingual Dutch households; of the infants that remained in the analysis, n=2 in the Known condition, n=6 in the Unknown condition, and n=6 in the Nonsense condition had a non-Dutch speaking parent.

Infants were included in analysis if they were typically-developing, born full-term (no more than 3 weeks preterm[2]), had good vision, and did not have more than one parent with diagnosed dyslexia. Further information about exclusions are detailed in the Data preparation section.

Infants were randomly assigned to one of the three conditions and to one of the two familiarization grammars (counterbalanced for sex).

3.3.2 Stimuli

The items corresponding to the Known and Unknown conditions were selected based on the Lexilijst Nederlands (Schlichting & Lutje Spelberg, 2009), a list of words and the percentage of children who recognize those words at various ages between 15 and 27 months of age.

The stimuli for the Known and Unknown conditions were color stock photographs from a collection assembled by Caroline Junge (Junge, Cutler, &

[2]According to the protocol of the Many Babies Consortium (unpublished protocol).

Hagoort, 2012). When this set did not contain one of the images needed, we found high quality creative commons images that matched visually in terms of quality and white background with those in the original set.

The visual stimuli for the Nonsense condition were created using images from the NOUN database (Horst & Hout, 2016).

Familiarization items in the Known condition were: X category: *apple, shoe, bed, ball*; Y category: *book, jacket, car, bath*. Test items were: X items: *bike, chair*; Y items: *banana, sock*.

Familiarization items in the Unknown condition were: X category: *leaf, cup, umbrella, plant*; Y category: *candle, pencil, chalk, balloon*. Test items were: X category: *helicopter, carrot*; Y category: *lego, chocolate*.

Items in the nonsense items were also divided into an X and a Y category, but are not possible to name. They have been given item numbers by Horst and Hout (2016) and were divided as follows: X category: *2011, 2015, 2006, 2017*, Y category: *2021, 2005, 2053, 2028*. Test items were: X category: *2025, 2023*; Y category: *2013, 2063*. A sample of the triads created are presented in Figure 3.1 and all individual items used are shown with label in the appendix Figure A1.

Pictures in all conditions were selected to differ visually, in terms of color, shape, and/or orientation when combined.

3.3.3 Procedure

Infants were tested in a Preferential Looking paradigm similar to that used in Saffran et al. (2007). During the experiment, infants sat on their caregiver's lap in a sound-attenuated booth. The caregiver wore glasses that were taped over with black duct tape to ensure they were blind to the visual stimuli. This blinding precaution was intended to help reduce caregiver interference during the trials while still allowing them peripheral vision such that they could still react to their child's needs if necessary.

The experiment consisted of a familiarization phase of fixed length[3], and a test phase. An examiner outside of the sound-attenuated booth followed the infant's behavior on a monitor and controlled the experiment accordingly. The experiment began with an attention grabber composed of soft purple-toned circles that moved slowly around the screen. A soft harp chord accompanied the visual stimulus. When infants focused on the attention grabber for more than two continuous seconds, the familiarization phase started automatically. The familiarization phase consisted of four blocks of 34 seconds each separated by attention grabbers[4]. Each familiarization block consisted of all 16 triads played once. Across the four blocks, infants were familiarized with the 16 triads played 4 times in different randomizations, resulting in 64 total triad exposures.

[3]Note that Saffran et al. 2007 used an infant-controlled Habituation procedure.

[4]The familiarization phase was blocked with attention grabbers in between because pilot testing indicated that, unlike in other experiments in our lab with auditory stimuli, infants became fussy very quickly in the sound-proof booth with visual stimuli only.

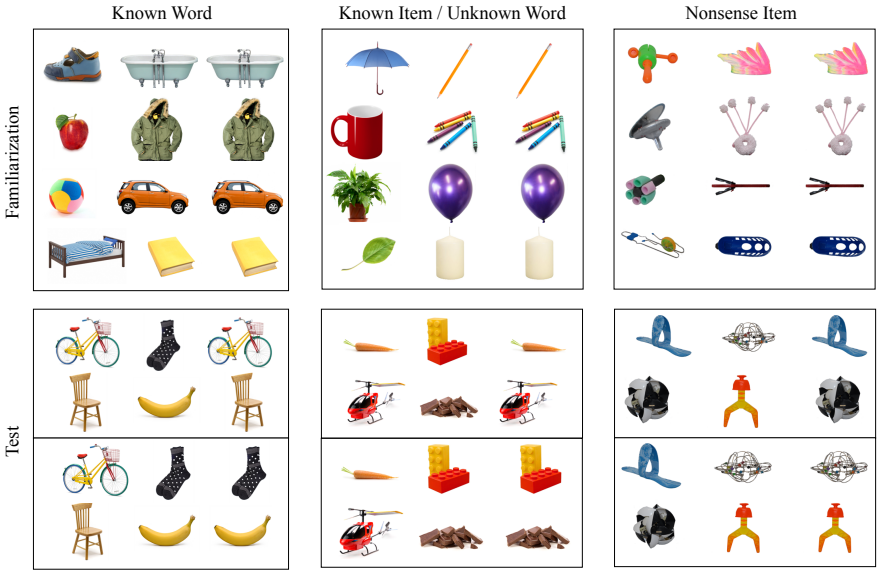


Figure 3.1: Examples of familiarization triads, and the test triads used in each of the three conditions. Appendix A displays all labeled items for all conditions.

The triads appeared on the screen sequentially and cumulatively in a manner replicating that described in Saffran et al. (2007). The first item appeared on the left of the screen for 330 ms, followed by the second item which appeared in the middle of the screen for 330 ms together with the first item, followed by the third picture which appeared on the right of the screen for a total 830 ms along with the previous two pictures. Thus each trial lasted 1.99 seconds. Triads were separated by a white screen that appeared for 500 ms. Figure 3.2 displays the time-course of the triad.

Stimulus presentation during the familiarization phase was not contingent on infants' looking towards the screen. If the infant looked away during the familiarization phase, the stimulus continued to play until it reached the attention grabber, at which point the infant would be re-focused.

The familiarization was followed by an attention grabber, and subsequently by the test phase. Each test trial was composed of one triad of novel items repeated seven times in the same presentation manner as the familiarization triads. This resulted in a trial duration of maximally 15 seconds see Figure 3.2). Each test triad either followed a grammar that was consistent or inconsistent with the familiarization grammar. There were two consistent grammar trials

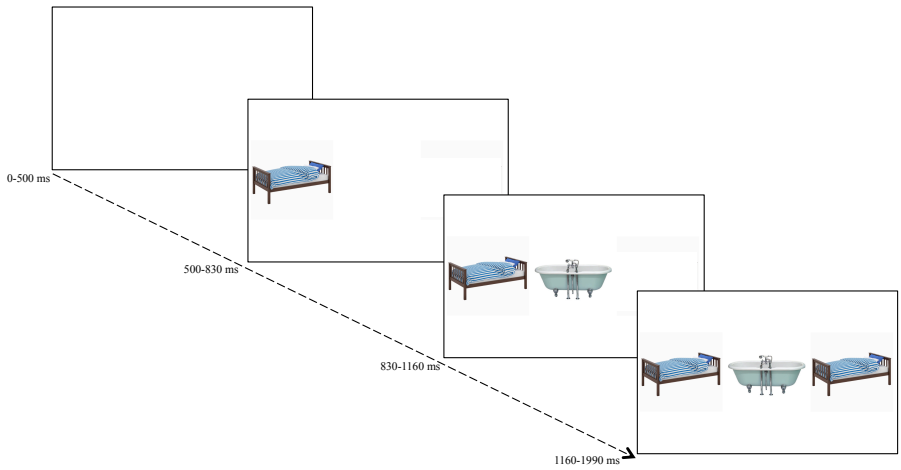


Figure 3.2: Time course of one triad. Triads were preceded by a blank screen presented for 500 ms, followed by the first item of the triad which was presented for 330 ms, the second item presented for 330 ms along with the first item, and finally, the third item, presented along with the previous two items for 830 ms total. During the familiarization phase, a triad would be followed by a different triad. During the test phase, a triad would repeat for a maximum of seven repetitions per trial.

and two inconsistent grammar triads. These were repeated three times each, resulting in three blocks of randomized test trials.

3.4 Results

3.4.1 Data pre-processing

Infants were excluded from analysis for the following reasons: having fewer than six test trials due to extreme fussiness or due to under- or over-looking (as described below; n=10; Known condition = 3, Unknown condition = 2, Nonsense condition = 2), for experimenter and technical errors (n=11; Known condition = 4, Unknown condition = 5, Nonsense condition = 2), for being born preterm (n=7; Known condition = 6, Unknown condition = 1), for having both parents diagnosed with dyslexia (n=1, Known condition), for falling asleep during the experiment (n=1, Known condition), and for standing during the experiment such that the face was not visible to the camera (n=1, Nonsense condition).

Infants' looking behavior was coded offline by hand using ELAN (*ELAN Version 5.0.0*, 2017). Behavior was coded as either looking towards the stimuli or not. A look indicated that the infants' eyes were directed towards the screen. Infants were not considered to be looking when their eyes were directed more than 45° away from the center of the screen in any direction.

Because a trial's visual stimulus lasted 15 seconds, this designated the maximum amount of time infants could look. LTs that reached or exceeded this maximum were removed to remove ceiling effects that can skew the data (see e.g., Bernard & Gervain, 2012; Gervain & Werker, 2013)[5]. This action resulted in the removal of 69 of the 564 data points (12%), 24 from the Known condition, 27 from the Unknown condition, and 18 from the Nonsense condition.

LTs of less that 1.16 seconds were also removed, as infants could not have seen the critical difference between grammars before this time (see e.g., Ferguson & Waxman, 2015)[6]. These types of trials were less frequent and constituted 5 data points in the Known condition, 11 in the Unknown condition, 18 in the Nonsense condition (6% of the data).

Infants were removed from analysis altogether if they had six or more trials out of the total 12 removed. Nine infants who were removed for having too few trials were removed for too many long look trials (n=5 in the Known condition, n=2 in the Unknown condition, n=2 in the Nonsense condition).

After these removals, we were left with 47 infants across the three conditions. The final number of participants whose data was analyzed was as follows: Known condition n=13 (7 girls); Unknown condition n=17 (7 girls); Nonsense condition n=17 (8 girls).

Finally, in keeping with the suggestions of Csibra, Hernik, Mascaro, Tatone, and Lengyel (2015) and the standardized protocols proposed by the Many Babies Consortium[7], all LTs were log base-10 transformed (the transformed dependent variable henceforth referred to as LogLT) before analysis.

3.4.2 Overall analysis

A Repeated Measures ANOVA was conducted in JASP (JASP Team, 2017) with the within-subject factor of Consistency (of the test grammar with the Familiarization grammar - either Consistent or Inconsistent) and Block[8] (1, 2, and 3) and the between-subject factors Condition (Known, Unknown, Nonsense), Familiarization Grammar (XYX and XYY), and Sex (male, female).

The overall analysis revealed a significant main effect of within-subject factor Block with a medium effect size [$F(2,46)=9.254$, $p<0.001$, $\eta_p^2=0.287$],

[5]Note that in chapter 2, analysis is done with and without these times. The decision was made in this chapter to only include LTs up to 15 seconds long because the looking behavior measured in the Preferential Looking Procedure used here was a less active task than the looking behavior measured in the Headturn Preference Procedure. This difference in task would allow infants to stare more easily in the former than the latter paradigm. For this reason, we considered that this procedure required a stricter criterion.

[6]These authors removed LTs below 2.5 seconds but we considered this criterion as too strict, since the crucial difference between the grammars in our experiment, as in theirs, comes at 1.16 seconds (see Figure 3.2).

[7]Unpublished protocol.

[8]Averaging within a block, rather than using each individual trial as a datapoint, mitigated the lack of robustness of this type of model to missing values, of which there were too many to use individual trials.

and a significant three-way interaction with a medium effect size between Block, Condition, and Sex [$F(4,46)=3.592$, $p=0.012$, $\eta_p^2=0.238$] (see Figure 3.3).

In addition, a marginally non-significant interaction between Condition and Sex, with a medium effect [$F(1,23)=3.216$, $p=0.059$, $\eta_p^2=0.219$] was found. These results indicated that infants' looking times differed throughout the experiment and that that difference in looking throughout the experiment differed between boys and girls, but also between the three conditions (see Figure 3.3).

Crucially, however, there was no main effect of, or interactions with, factor Consistency (the factor which would have indicated if learning took place), and no interaction of Consistency with Condition, which would have indicated a difference in learning between the different conditions (see Figure 3.4). In addition, there was no interaction between Consistency and Familiarization grammar, an interaction which would have shown a general preference for one pattern or another. Furthermore, Familiarization grammar did not yield a significant main effect or a significant interaction, indicating that the grammar that infants were exposed to during familiarization did not make an overall difference in their looking behavior at test. Thus, infants' looking did differ but was not reflective of learning or any preference with respect to the patterns they were presented with.

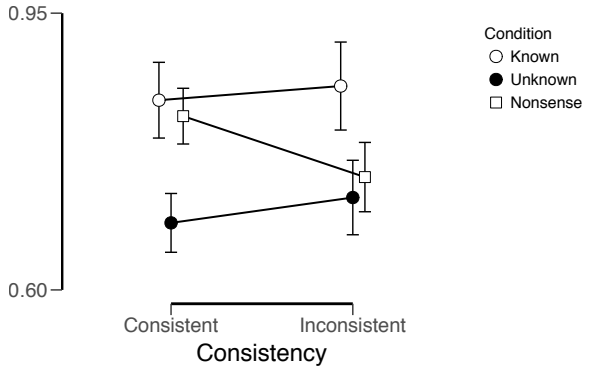
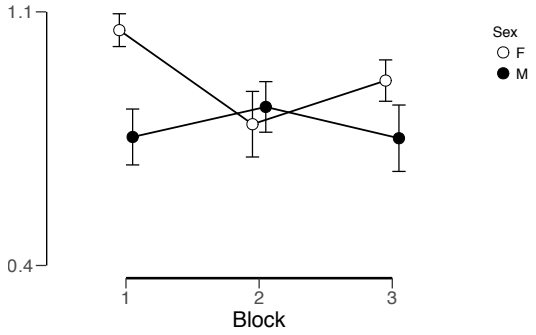


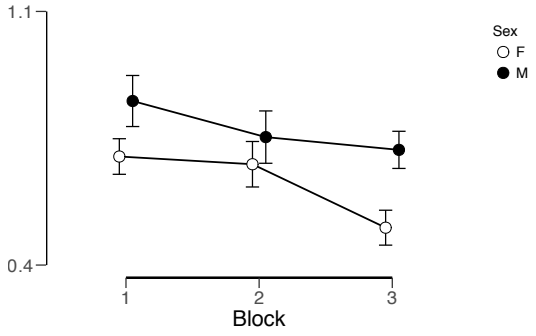
Figure 3.4: LogLTs towards consistent and inconsistent test items per experimental condition.

Bayesian analysis

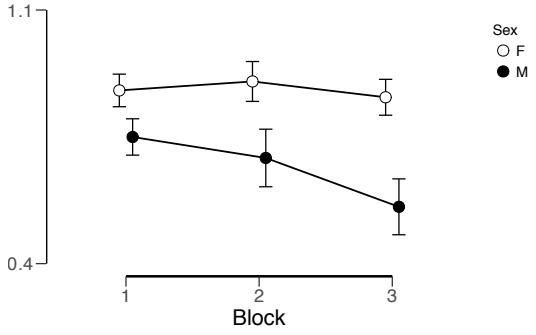
In order to confidently state that the null hypothesis is likely to be true, Bayesian paired t-tests were conducted between the looking times to Consistent vs. Inconsistent grammars using the data from all three conditions. This overall analysis reveals moderate evidence for the null hypothesis in Block 1 ($BF_{01}=5.796$), Block 2 ($BF_{01}=5.326$), and Block 3 ($BF_{01}=4.597$). When analyzing the grand average (unblocked) of LogLTs, the evidence is also moderate for the null hypothesis ($BF_{01}=6.315$).



(a) Known Condition



(b) Unknown Condition



(c) Nonsense Condition

Figure 3.3: LogLTs across the three test blocks in the three different conditions divided by sex.

Exploration of covariates

We investigated whether two additional factors might play a role in infants' looking behavior: NCDI score and LT during familiarization (LTFam). NCDI is taken into account because the goal of this experiment was to understand whether lexical knowledge of items plays a role in learning how they are structured. There is a possibility that a larger vocabulary in general might allow infants to more quickly shift their focus from individual items to their underlying rules. In addition, the LT during familiarization may play a role, as it may be a measure of interest and focus, or on the other hand, a measure of struggling to learn.

When including these two factors as covariates, a main effect of LTFam [$F(1,21)=9.418$, $p=0.006$] was found. The three-way interaction of Block, Condition, and Sex found in the original model remained [$F(4,44)=2.986$, $p=0.029$]. However, NCDI score was not significant.

3.4.3 Analysis per condition

Known Condition

In the Known condition, there were significant interactions between Block and Sex [$F(2,12)=10.625$, $p=0.002$, $\eta_p^2=0.639$] and Block and Familiarization [$F(2,12)=4.928$, $p=0.027$, $\eta_p^2=0.451$], as well as a strong main effect of Sex $F(1,6)=9.418$, $p=0.022$, $\eta_p^2=0.611$].

As in the larger model, no evidence of learning was found, nor a general preference for one type of grammar or another. The inclusion of the covariates LTFam and NCDI score did not reveal any main effects or interactions with these covariates, indicating that they did not have an impact on infants' learning.

The significant interaction between Block and Familiarization grammar indicates that infants familiarized with the XYX grammar did not lose interest throughout the course of the experiment, while infants familiarized with XYY had the same looking time in the first block, but looking time decreased in the second and third blocks (see Figure 3.5). The interaction of Block and Sex also showed that while boys mostly kept looking steadily throughout the course of the experiment, girls looked significantly more in the first block than boys, but looked significantly less in the second and third block (see Figure 3.6).

These significant interactions with Block indicate that behavior in Block 1 may be different from behavior in Blocks 2 and 3. We thus removed the within-subject factor of Block and conducted an exploratory analysis using only the LogLTs during Block 1 to investigate whether a learning effect showed up immediately after familiarization.

In Block 1, there was a significant interaction between Consistency and Sex [$F(1,8)=7.296$, $p=0.027$, $\eta_p^2=0.477$], a significant main effect of Sex, and a significant main effect of Familiarization Grammar [$F(1,8)=15.368$, $p=0.004$,

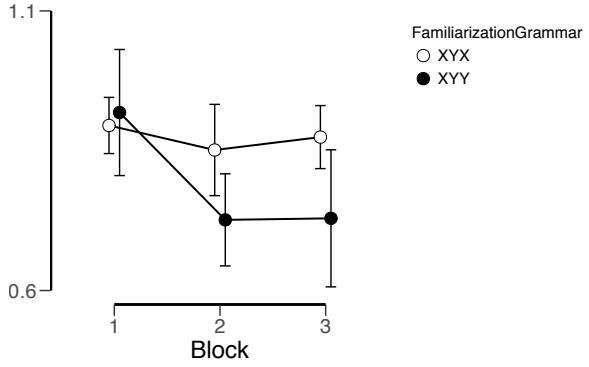


Figure 3.5: LogLTs over time of infants in the Known condition familiarized with the XYX vs. the XYY grammars indicating a decrease in looking over time in infants trained with XYY but not XYX.

η_p^2=0.658). While girls looked significantly longer than boys overall in the first block (M_{diff}=0.299, p_{bonf}<0.001), the interaction between Sex and Consistency revealed that it was boys who showed evidence of learning, looking significantly longer to the inconsistent grammar (M=0.893, SE=0.037) than to the consistent (M=0.644, SE=0.049), while girls showed a very slight preference for consistency (M=1.087, SE=0.041) over inconsistency (M=1.048, SE=0.031).

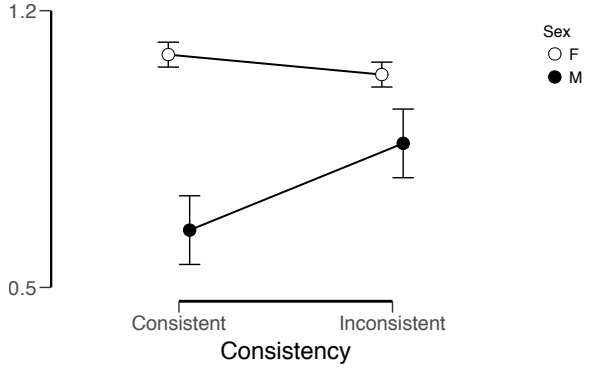


Figure 3.6: LogLTs of infants to Consistent vs. Inconsistent grammars in the Known condition in in Block 1 only.

This exploration indicated that a learning effect may actually be present but that over the course of the experiment, as interest decreases and memory of the familiarization grammar may fade, this effect is obscured. Unfortunately, because of our stringent inclusion criteria, we have little power in this condition,

and strong conclusions should be avoided. However, the results here do show that learning may have been occurring in this specific condition and should be further tested in light of the differences with the other conditions.

Unknown Condition

In the Unknown condition, the analysis revealed a main effect of Block [$F(2,16)=5.987$, $p=0.011$] and no other main effects or interactions. Post-hoc analyses revealed that Block 1 differed from Block 3 ($p_{bonf}=0.011$) but not from Block 2 ($p_{bonf}=0.23$). Block 2 and 3 did not differ from each other. This result indicated a general decline in interest of the infants over time. Inclusions of NCDI score as covariate did not reveal a significant main effect of NCDI score, nor any interactions, and was thus ignored. Inclusion of LTFam however did yield a significant effect of LTFam [$F(1,7)=15.323$, $p=0.006$]. Including this covariate rendered the effect of Block non-significant. These results taken together indicate that infants who were long lookers during the familiarization remained long lookers during the test, but that the test on its own had no effect on their looking time in general.

In this condition, we found no evidence of learning and no evidence of a general preference for one grammar over another.

Nonsense Condition

In the Nonsense condition, there were no significant main effects and no significant interactions between Consistency, Block, Sex, or Familiarization Grammar indicating that none of these factors helped explain the variation in this data set.

Thus, to test the effect of other potential factors as above, NCDI score was added to the model as a covariate. The covariate was significant overall [$F(1,8)=7.132$, $p=0.028$, $\eta_p^2=0.471$], but it did not play a role in behavior towards these test grammars[9]. In addition, with the addition of this covariate, the main effect of Sex was also significant [$F(1,8)=9.855$, $p=0.014$, $\eta_p^2=0.552$]. In a Bonferroni post-hoc analysis, girls looked significantly longer than boys ($M_{diff}=0.234$, $p_{bonf}=0.042$). LTFam in this case was not significant as a covariate.

As in the previous conditions, there was no sign of learning from the experiment, nor of a general preference for the one type of grammar over the other, and no evidence that the familiarization phase had an impact on infants' learning. In addition, in this condition, there was no effect of block either, showing that in this condition infants did not decrease their interest over time. These

[9] To test whether higher vocabulary scores correlated with (better) learning, the factor of Vocabulary was added (High and Low based on a median split at 31 words). However, this factor did not prove to be significant, nor did it produce any interactions. Thus while their NCDI score did seem to have an effect on infants' overall looking (LogLTs), it did not play a role in preferences with respect to the test grammars.

results indicate that any variation in the data in this condition resulted from individual differences, differences related to sex, and more general knowledge such as number of words known.

3.5 General Discussion

In this experiment, we set out to investigate whether lexical knowledge provides the basis for learning about structure. We hypothesized that infants would be able to learn from a brief exposure period to visually-presented grammars XYX or XYY when they were composed of items of which they knew the name (Known condition) but not when they did not or could not know the name of the item (Unknown condition and Nonsense conditions).

Our results showed that a very specific subset of the infants were able to generalize, namely boys in the Known condition, familiarized with XYY, in Block 1. These are admittedly weak results. However, the fact that these boys looked longer to the inconsistent XYX grammar than to the consistent XYY grammar indicates that learning from the familiarization phase was completed by the time the test phase started (looking to novel items typically indicates a completion of learning during familiarization). On the other hand, for the girls, learning during the familiarization phase did not seem to be completed, as seen by their by their longer looking times overall during the test coupled with their lack of discrimination between the grammars. While these results indicate that some learning seems to be possible, they are not robust in terms of power, and generalization of interpretation should be done with caution.

In addition, our findings in the Unknown and Nonsense conditions point to variation as a result of individual differences. In the Unknown condition, infants varied on whether they were "long lookers" or "short lookers" overall. Those infants who maintained their looking during the familiarization continued to maintain their looking during the test, although this maintenance in attention did not result in any preferences with respect to the test items. In the Nonsense condition, boys and girls differed, and this was influenced by whether they had a high or a low vocabulary. However, neither in the Nonsense condition, nor in the Unknown condition did these differences translate into preferences with respect to test items. Taken together, the results of the three conditions give us an indication that while in the Unknown and Nonsense condition, data varied with respect to participants' individual differences, in the Known condition, data varied more with respect to the experimental manipulation.

These differences between the Known condition and the other two conditions indicate that lexical knowledge could in fact be playing a role in learning about the pattern. This may be because the lexical knowledge of the items helps the infants to analyze the stimuli, as they can potentially name the items covertly as they appear on screen. Lexical knowledge could thus be helping to direct infants' attention to the pattern, to categorize the stimuli, and to memorize the regularity during familiarization.

3.5.1 Conclusion and future work

Our results, although not robust, indicate that lexical knowledge of items presented in a rule learning task could aid in pattern generalization. The experiments must be replicated, and extended upon, to verify that the effect found with the specific set of participants here was not a chance finding. Nevertheless, the results of this study help us direct our attention for future work.

What we can say with more certainty is that in the conditions in which familiarization triads consisted of familiar items with unknown names, or of nonsense items, there was no evidence that infants learned an underlying XYX or XYY rule. Thus, rather than the mere familiarity of items, the crucial facilitating factor in learning rules with visual stimuli may be the availability of a lexical specification of the pictured item. If learning in Saffran et al. (2007) simply occurred because of the familiarity of the pictured items, infants in our Known and Unknown conditions, which both contained items that were familiar to them, should have performed comparably. Instead, we found that infants had much longer looking times in the Known condition, both during familiarization and during test, than in the other conditions. We found an indication of learning only in this Known condition, although both exposure grammar (XYY) and sex (male) of the participant played a role as well.

In order to confirm the findings presented here, more infants should be tested particularly in the Known condition. In addition, several further considerations should be mentioned with respect to follow-up work.

The role of repetition

The only significant results we found were in boys trained with the XYY grammar. In much of the literature, and also in our own results presented in chapter 2, we find that repetition is a highly facilitating factor. Gómez, Gerken, and Schvaneveldt (2000) argue that repetition is the basis of transfer in AGL tasks. In their work with adults, they found that participants could only generalize from a familiarized finite state grammar to novel input on the basis of the presence of repetition. Experiments looking specifically at Marcus-type grammar learning with both infants (S. P. Johnson et al., 2009) and adults (Endress et al., 2007; Chen et al., 2015) have shown that repetition is a highly salient cue for rule learning. Repetitions, particularly in edge positions, have been identified as "perceptual or memory primitives" that form one of the principal mechanisms by which we are able to attend to and learn artificial grammars (Mehler et al., 2008; Endress et al., 2009). It may thus be that learning rules containing adjacent repetitions is the default, and that extra support is necessary to move beyond this perceptual bias.

In our Known condition, the fact that infants lost interest over time when being exposed to the XYY grammar, but not to the XYX grammar, can be interpreted as evidence that some level of familiarization with the repetition pattern had taken place. In addition, just as in S. P. Johnson et al. (2009)'s

findings with eight- and 11-month-old infants, boys in our Known condition were unable to generalize after exposure to the XYX pattern but showed evidence of generalization from the XYY grammar. Our findings thus add to the claim that the presence of the repetition during familiarization is a facilitating factor.

Differences by sex

While we specifically counterbalanced the number of boys and girls per experimental manipulation to make sure there would not be a confound of sex, with the numerous post-test removals of data points and participants, we did not always have an equal number of data points per factor for each sex. However, there is a potential difference between sexes worth exploring in future (visual) rule learning studies. Literature on differences between sexes in infant studies is scarce, making it difficult to integrate our results with previous findings. Nevertheless there is some evidence from linguistic tasks that young infants may already show a difference in preference and processing of patterns depending on their sex. For example, Shi (2007) showed that eight-month-old girls preferred frequent function words, while boys preferred infrequent ones. Ter Haar and Levelt (2018) found that 12-month old boys, but not girls, tended to prefer sets of frequent over infrequent CV syllables.

Differences between the sexes have also been found in previous rule learning studies. In an ERP study, Mueller, Friederici, and Mannel (2012) found that three-month-old girls showed a mismatch negativity to stimuli violating a rule, while boys showed a positivity. While both detected the mismatch between the standard and deviant, the authors concluded that the girls, whose response was similar to adults', were more mature at this age with respect to their processing. Lany and Gómez (2008) also found that at 12-months-old, girls were better than boys at learning adjacent dependencies between word categories and generalizing the dependencies to non-adjacent positions. These studies indicate that there are differences between sexes in their processing of structured information that require further investigation.

Role of variety

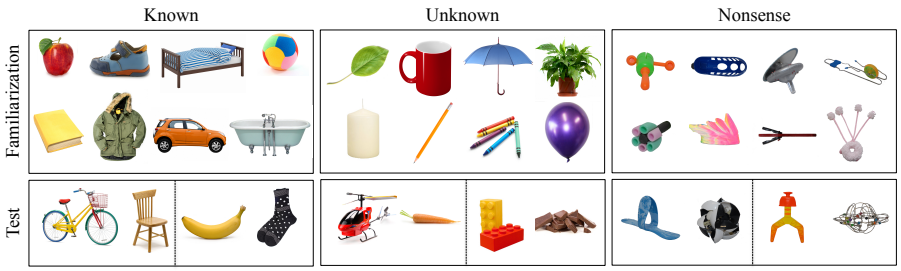
Visual inspection of our stimuli (see Figures 3.1, A1) and those found in Saffran et al. (2007) reveal that, compared to the minimally different pictures of dogs and cats in the triads used in Saffran et al., the X and Y images we used within a triad were very different from each other. The minimal differences in Saffran's stimuli may have played an important role in infants' ability to learn rules. The X and Y pictures in our triads varied on multiple dimensions: shape, color, orientation, and identity value (i.e., items were not different versions of the same object). While this may have been interesting for infants overall, it may have impeded them from discovering an underlying rule. We know from previous research that there is a "sweet spot" of complexity with respect to the

stimuli presented to infants and their ability to learn at various levels (e.g., "The Goldilocks Effect" of Kidd, Piantadosi, & Aslin, 2010); learning cannot occur if stimuli are too complex, nor if they are too simple and are not informative enough. The uniformity of the triads in Saffran et al. (2007), consisting only of very similar dogs (or cats), might have helped the infants to understand that the triad formed some sort of unity (according to the Gestalt principles of perception and similarity), within which there was a pattern of same and different (XXY or XYX; Newport & Aslin, 2004). The triads in our study, on the other hand, did not form a clear unity, or Gestalt, in any of the conditions. If both lexical knowledge and Gestalt principles of similarity are facilitating factors, this would account for the stronger learning effects in Saffran et al. (2007). Further research looking at the effect of minimal differences between items forming a rule, and the interaction with lexical knowledge is the necessary next step in understanding the success of infants in Saffran et al. (2007).

3.6 Appendices

3.6.1 Appendix A

Figure A1: All items used during the familiarization and test phases in the Known, Unknown, and Nonsense Conditions. Top rows of each Familiarization box were used as X category items, and bottom rows were used as Y category items. Test items boxes are divided by a dotted line, the left of which were X category items, the right of which were Y category items. Nonsense familiarization items were: Top row – 2011, 2015, 2006, 2017, Bottom row – 2021, 2005, 2053, 2028. Nonsense test items were: 2025, 2023, 2013, 2063.



3.7 Acknowledgements

Thank you to Claire van Leeuwen and Marjolein Maljaars for your help in running the experiments, and to Marijn van 't Veer for your help in scheduling the appointments.

CHAPTER 4

Patterns in infant babbling

This chapter appears as Geambaşu, A., Scheel, M., & Levelt, C.C. (2016) Cross-linguistic Patterns in Infant Babbling. *Proceedings of the 40th annual Boston University Conference on Language Development*, ed. Jennifer Scott and Deb Waughtal, 155-168. Somerville, MA: Cascadilla Press.

In this chapter, the goal was to better understand what types of patterns infants produce in the age range at which they perform perceptual rule learning tasks. A better understanding of how babbling progresses structurally can shed light on successes and failures in perceptual tasks.

4.1 Abstract

Previous literature has indicated that infants' babbling is characterized by an initial phase of reduplication to a later phase of variegation. However, a systematic analysis of patterns in infant babbling across languages and age groups has never been undertaken. We thus analyzed the languages available with phonetic transcriptions in the Phon database (Rose & MacWhinney, 2014). We investigated more than 60,000 two-, three-, and four-syllable utterances from 5- to 24-month-old infants across nine languages. We found that already from the earliest age group, infants produce variegated patterns significantly more than any other possible patterns, and that this does not change over development. In addition, infants of 18-24 months show a drastically increased rate of variegation (approx. 20% increase), likely a result of the ontogeny of the lexicon. Implications both for theories of development and for experimental work will

be discussed.

4.2 Introduction

Infants begin to vocalize very soon after birth, and they begin to babble about six months after they are born (Oller et al., 1980). The babbling stage is distinct from the previous phase of vocalizations in that sounds – or gestures in infants acquiring sign language – are now clearly organized in a syllabic structure. As such, these utterances are the infant's first linguistic productions.

In the works of Stark (1980) and Oller et al. (1980), two stages were identified within the babbling phase. Babies start with reduplicated babbling when they are six to eight months old, and progress into 'variegated' (Oller et al., 1980) or 'non-reduplicated' (Stark, 1980) babbling at 10 to 12 months. Work by Koopmans-Van Beinum and van der Stelt (1986) outlines a similar line of development, with reduplicated babbling beginning at six months and lasting up until at least 12 months. They do not identify a specific non-reduplicated stage during this period. In addition, Roug et al. (1989) also identified babbling stages similar to those proposed by Oller and Stark, with reduplicated (consonant) babbling beginning at seven months, and variegated babbling beginning at 12 months. The stages identified by these researchers differ only slightly. Where they crucially converge is on the consensus that infants begin their babbling at around six to eight months old, that they begin with reduplicated utterances, and that they transition into producing variegated utterances at around 10 to 12 months. The existence of these two stages has been disseminated in introductory linguistics textbooks for years (e.g., Hoff, 2008). Yet, the data on which these particular stages have been based came from relatively small-scale studies and primarily from English-acquiring infants.

In the past decades, some studies have focused on finding out whether the proposed stages still hold when data from a larger number of infants is taken into account, and whether they take the same form across languages. Smith, Brown-Sweeney, and Stoel-Gammon (1989) in one notable example followed 10 English-learning infants from six to 18 months old, at three four-month intervals. They found that while reduplicated utterances occurred more frequently in the first two age groups (six to nine months and 10 to 13 months respectively), both reduplicated and variegated utterances co-occurred across all three age groups. Crucially they did not find a change with respect to the frequency of reduplicated versus variegated babbling at the time when this change was expected according to the above-mentioned literature, namely between the first and second age groups. More recently, a study comparing songbird and infant babbling by Lipkind et al. (2013) analyzed data of nine English-acquiring infants from the Davis corpus (Davis & MacNeilage, 1995) and showed that there was no difference between the ages of 8- through 29-months old in American English-learning infants with respect to the amount of reduplicated babbling found. Thus they found no discrete stage of reduplication followed by one of

variegation. Instead, what they found was that infants and birds both tended to reduplicate newly-acquired sounds more than other sounds.

While Lipkind and colleagues present a crucial first step in a larger-scale approach by making use of the data available in the CHILDES database (MacWhinney, 2000), a wealth of previously-collected data that can give us an even better understanding of babbling has remained unexplored up until now. For the present study, we collected and systematically analyzed available corpus data from different languages in order to finally determine unequivocally what the universal developmental babbling patterns are, and to assess language-specific influences on babbling. With this work, we thus present the first large-scale, cross-linguistic survey of the babbling patterns of infants.

Our hypothesis was that if there are distinct developmental stages with respect to babbling patterns (Oller et al., 1980; Stark, 1980; see de Boysson-Bardies, 2001; Hoff, 2008; Vihman, 1996, etc. for reviews), we should see predominantly reduplicated babbling in the period of seven to ten months and a switch to predominantly variegated babbling in the period of ten to 12 months. An alternative hypothesis, based on the findings of Koopmans-Van Beinum and van der Stelt (1986) and of Roug et al. (1989), was that we should see predominantly reduplicated babbling throughout the entire period of seven through 12 months old. This would indicate that variegated babbling only truly begins later, with the ontogeny of the lexicon. The aim of this work is thus to pinpoint if and where a transition from one stage to another may occur. In addition, by conducting a cross-linguistic analysis, our aim is to determine which aspects of babbling development are universal, and to identify language-specific variation.

4.3 Language and Corpora Selection

The present study was carried out on datasets of child utterances from different languages, made available through the PhonBank corpus (Rose & MacWhinney, 2014). The datasets that were suited for our purpose were collected and transcribed by seven individual principle investigators for the purpose of their own varied research goals. The transcriptions have since been made available by these researchers to be used and analyzed by others with the aid of the Phon software (Rose et al., 2006; Rose & MacWhinney, 2014).

From all the available datasets (n=33, from 12 different languages) represented in PhonBank, we selected only those datasets that included phonetic transcriptions in the International Phonetic Alphabet (IPA) and that included child utterances recorded between our target age range of seven through 12 months. The lower limit of seven months was chosen, as this was the first age group for which phonetically transcribed data in at least one language was available, and, from a theoretical point of view, it is the lower end of the age range at which infants begin to babble. The upper limit was chosen because it is the age at which we expect variegated babbling to either begin to appear or even to overtake reduplicated babbling.

Datasets from eight different languages – Arabic (Kern, Davis, & Zink, 2009), Dutch (Kern et al., 2009), English (Providence corpus, Demuth, Culbertson, & Alter, 2006; Stanford corpus, de Boysson-Bardies & Vihman, 1991), French (Kern corpus, Kern et al., 2009; Lyon corpus, Demuth & Tremblay, 2008; Stanford corpus, de Boysson-Bardies & Vihman, 1991), Japanese (Ota corpus, Ota, 2003; Stanford corpus, de Boysson-Bardies & Vihman, 1991), Portuguese (Freitas, 1997), Romanian (Kern et al., 2009), and Swedish (de Boysson-Bardies & Vihman, 1991) – fit the criteria described above. Table 4.1 lists the languages, the corpora from which those languages were extracted, and the number of infants represented at each age group between seven and 12 months (per month, per corpus). Note that some languages are represented by more than one corpus, and that some languages are not represented at all ages.

Table 4.1: Languages studied, the names of the corpora from which the data was taken, the country of recording, and the number of infants per age group represented in each language.

			Age in Months					
Language	**Corpus**	**Country of Recording**	**7**	**8**	**9**	**10**	**11**	**12**
Arabic	Kern	Tunisia	0	3	3	4	4	3
Dutch	Zink	Belgium	0	4	4	4	4	4
English	Providence Stanford	USA	4	6	10	17	18	19
French	Kern Lyon Stanford	France	1	2	5	8	9	12
Japanese	Ota Stanford	USA	0	0	1	1	0	3
Portuguese	Freitas	Portugal	0	0	0	1	5	6
Romanian	Kern	Romania	1	4	4	4	4	3
Swedish	Stanford	Sweden	0	0	5	2	2	2

The data included recordings from monolingual households. In most cases, the infants were growing up in their country of origin. The Japanese infants were recorded in the United States (San Francisco in the case of the Stanford corpus, Washington DC in the case of the Ota corpus), but were raised in monolingual Japanese households.

4.4 Corpus Data Collection

In order to identify and analyze reduplicated and variegated patterns, the data had to be processed in such a way that each syllable was identified, labeled, and compared to the other syllables in the utterance. In this way syllables could be identified as the same (reduplicate) or different (variegated). Our interest was

not in the kind of syllables that were being reduplicated or used in variegated sequences, but in the extent to which reduplication and variegation was present. Thus after syllables in each utterance were identified, they were labeled in an abstract way (assigned a unique letter within each utterance). Then, each type of pattern was tallied for analysis. This process was done with the help of two scripts that will be described in more detail below.

The first step that was necessary to identify the patterns in infant utterances was syllabification. This was done with Phon's built-in syllabifier. Syllabification was applied to all records of each language. For all languages except for two, a language-specific syllabifier was available that syllabifies according to the rules of the language. The exceptions were Romanian and Japanese. For these languages, the English syllabifier was used, and complex utterances (eg., those with consonant clusters and geminates, longer utterances, etc.) were manually inspected. No odd or incorrect syllabification was found.

For the next steps, namely abstraction from the phonetic content of the syllable and comparison of two subsequent syllables, two scripts were written by Greg Hedlund and Yvan Rose and added to the Online Scripts Library in Phon (Phon 2.1, Rose et al., 2006). The first script, called SyllablePatternListing (Hedlund & Rose, 2015, https://www.phon.ca/phontrac/wiki/scriptlibrary), identifies the first syllable of an utterance, assigns it a letter (A), moves on to the second syllable, and performs an identity comparison with the previous syllable. If the first and second syllables are identical, the second syllable is also labeled A, yielding a reduplicated AA pattern. If the first and second syllables are not identical, the second syllable is labeled B, yielding a variegated AB pattern. For utterances of longer length, the identity matching is performed on each syllable as described above, and the comparison made with each of the previous syllables. Table 4.2 shows the possible patterns identified by this script. A second script named SyllablePatternSummary, tallied each of the patterns identified per infant.

4.5 Results

Our data collected from Phon consisted of 42,348 total utterances of one-, two-, three- and four-syllable utterances from a total of 64 infants. Utterances longer than four syllables were uncommon, and, from a practical point of view, beyond the scope of the current exploratory work. Of the one- through four-syllable utterances, 47% were one-syllable utterances (1SU), which cannot inform us about syllable patterns. We thus focused our analysis primarily on two- (2SU, 38%), three- (3SU, 11%), and four-syllable utterances (4SU, 4%).

Figure 4.1 illustrates how frequently infants in each language produced each type of utterance, as well as the means across languages. Note that the longer the utterance, the less frequently it is represented across almost all languages, the exception being Portuguese, where 2SU (57.6%) are more frequent than 1SU (25.4%). In Dutch, English, French, and Romanian on the other hand,

Table 4.2: All possible combinations of syllables in two-, three-, and four-syllable utterances, categorized roughly as reduplicated or variegated by their predominant pattern. Bold pairs of syllables are reduplicated. Underlined pairs of syllables are variegated.

	Utterance length		
Utterance Type	Two-syllable	Three-syllable	Four-syllable
Complete Reduplication	**A A**	**A A A**	**A A A A**
Combination - 3 reduplications			
Initial		**A A** B	**A A A** B
Final		A **B B**	A **B B B**
Internal			**A A B B**
Combination - 2 reduplications			
Initial			**A A** B C
			A A B A
Final			A B **C C**
			A B **A A**
Internal			A **B B** C
			A **B B** A
Complete Variegated	A B	A B A	A B A B
		A B C	A B A C
			A B C B
			A B C A
			A B C D

1SU made up 60% and above of collected utterances.

In our primary analysis we will focus on the 2SU and 3SU. Because of their complexity (see Table 4.1), 4SU will be discussed separately.

Across languages and across our age groups, the distribution of 2SU was 18.7% reduplicated and 81.3% variegated. When all age groups were pooled together, Romanian had a higher instance of reduplicated 2SU than other languages (30.3%), while English had a notably low rate of reduplication (6.2%). Figure 2 shows the distribution of reduplicated (AA) and variegated (AB) utterances in each language.

As shown in Table 4.1, 3SU can be broken down into five different patterns: completely reduplicated (AAA), completely variegated (ABC), or containing some element of both reduplication and variegation (ABB, AAB, and ABA). Figure 4.3 presents these patterns separately per language, pooling together all age groups. It is immediately evident that the overwhelming majority of utterances fell into the completely variegated pattern (ABC). Across all languages, this pattern represented a mean of 65.5% of all three-syllable utterances. In English, 83.3% of 3SU were completely variegated, while only 3.2% were completely reduplicated. All languages showed at least 50% completely reduplicated utterances in the distribution of the patterns (Romanian was at 50%, while all

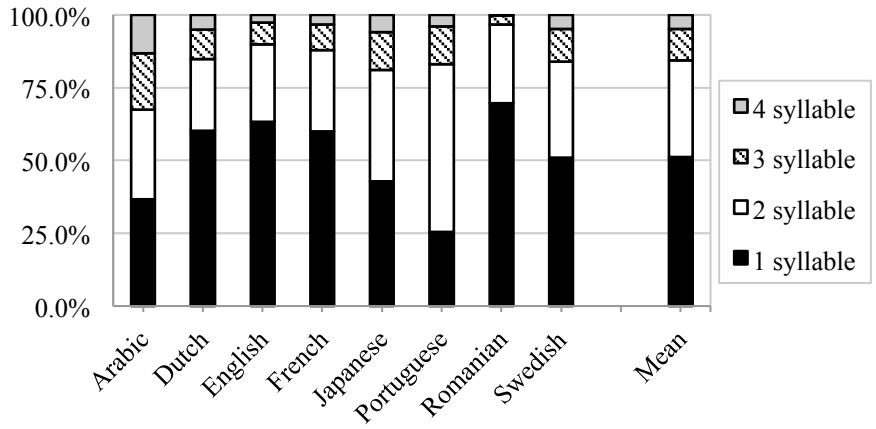


Figure 4.1: Percentage of syllables represented in each language, and the mean of all languages.

other languages were at approximately 65%).

The initial picture of the distribution of patterns across all age groups of 2SU and 3SU clearly shows that across languages, within the seven through 12-month-old range, infants actually tend to produce more variegation than reduplication. We now turn to answering our main question: can we differentiate an early stage with mostly reduplicated babbling and a late stage with mostly variegated babbling in our large, cross-linguistic dataset?

4.5.1 Analysis

Two-syllable utterances

Although our total dataset included 65 infants across eight languages, and all of these infants are presented in the figures for complete visual inspection (see Figure 4.4), for the purpose of a proper statistical analysis, some pruning of the data had to be done. Of the 65 infants, 14 individuals contributed only one month worth of data. We removed those infants who were not followed longitudinally in their respective studies, leaving a total of 51 infants. Of these 51, only six infants (12%) in three languages have data in the seven-month range. This month was thus also removed from analysis.

As the original recordings were collected by several different researchers for different research purposes, they also varied with respect to their frequency of collection. Some of the corpora contained multiple recording sessions per month while others included only one session. In order to make the data uniform with the respect to the number of data points per child, the tallies of multiple sessions within one month were averaged. In this way, each child had one datapoint per month, indicating how many utterances of each type he or she produced on

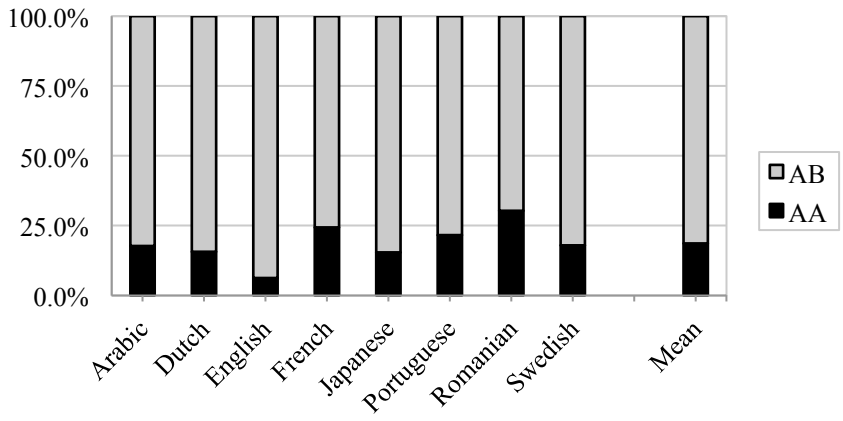


Figure 4.2: Distribution of two-syllable utterances across all ages following a reduplicated pattern (AA) and a variegated pattern (AB), in each language individually, and a mean of all languages.

average in that month.

This dataset was analyzed with a generalized estimating equation [GEE; geeglm in R, R Development Core Team, 2015). GEEs, as opposed to linear mixed models, do not attempt to model the individual variability in the data, and can thus tell us more about the mean performances. The analysis was run per pattern, with fixed effects of month and language. The factor of month was centered at 10 months to avoid correlation between estimates of the intercept and the slope. In addition, English was chosen as the reference language as this language had the most data.

On pattern AA, no interaction between month and language was found (χ^2=3.2, df=6, p=0.78), and no effect of month was found (χ^2= 0.07, df=1 p=0.79). An effect of language was found (χ^2=29, df=6, p<0.001). As no interaction was found, this factor was removed, and a second model was run. Here too we found no effect of month (χ^2=8.06, df=6, p=0.09), and an effect of language (χ^2=29, df=6, p<0.001). A Wald test reveals that Dutch and Romanian differed significantly from English (Wald=5.46, p=0.02 for Dutch; Wald=14, p<0.001 for Romanian). Both languages produced significantly more AA patterns than English (see Figure 2).

The same analysis was run on the AB pattern and while no interaction between month and language was found (χ^2=5.24, df=6, p=0.5), a significant effect of month (χ^2=4.25, df=1, p=0.04) and a significant effect of language (χ^2=17.26, df=6, p=0.008) were found. A second model with the interaction effect removed yielded identical effects for month and language respectively. Although significant effects of month and language were found, in the initial model, further inspection of the details with a Wald test revealed no effect of month (Wald=3.39, p=0.07), and no individual language differed from English.

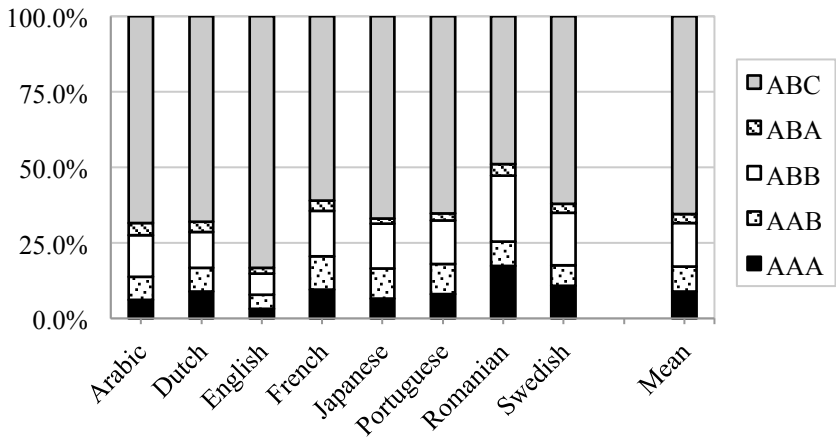


Figure 4.3: Distribution of three-syllable utterances following a totally reduplicated pattern (AAA), a totally variegated pattern (ABC) and patterns containing some intermediary level of reduplication (ABB, AAB, ABA).

Three-syllable utterances

All 3SU were analyzed in the same way as 2SU. Their development can be inspected visually in Figure 4.5.

For pattern AAA, no interaction of month and language was found (χ^2= 9.68, df=6, p=0.14) and no effect of month was found (χ^2=0.79, df=1, p=0.37. An effect of language was found (χ^2=12.57, df=6, p=0.05). A second model that excluded the interaction effect was run, confirming the presence of the language effect (χ^2=12.71, df=6, p=0.05). Wald tests reveal that only Dutch differed significantly from English with respect to this pattern (Wald=8.94, p=0.003), with higher rates of AAA overall.

In pattern AAB, there was a significant interaction between month and language (χ^2=14.5, df=6, p=0.025) and a significant effect of language (χ^2=115.8, df = 6, p<0.001). However, no effect of month was found (χ^2=0.1, df=1, p=0.772). Dutch (Wald=11.11, p<0.001) and Romanian (Wald=4.94, p=0.02) were significantly different from English for this pattern. In addition, this model also found an interaction between language and month in French (Wald=5.85, p=0.02) and Romanian (Wald=5.24, p=0.02), both showing less of this pattern over time. Particularly in Romanian, this pattern largely disappears after 10 months.

For pattern ABA, there was an interaction between month and language (χ^2=15.8, df=6, p=0.02), and an effect of language (χ^2=21.5, df=6, p=0.002), but no effect of month (χ^2=2.1, df=1, p=0.15). Only Portuguese differed significantly from English over time (Wald=8.48, p=0.004), with Portuguese babies producing this pattern at 11 months but not at 10 or 12 months.

For pattern ABB, there was no interaction between month and language

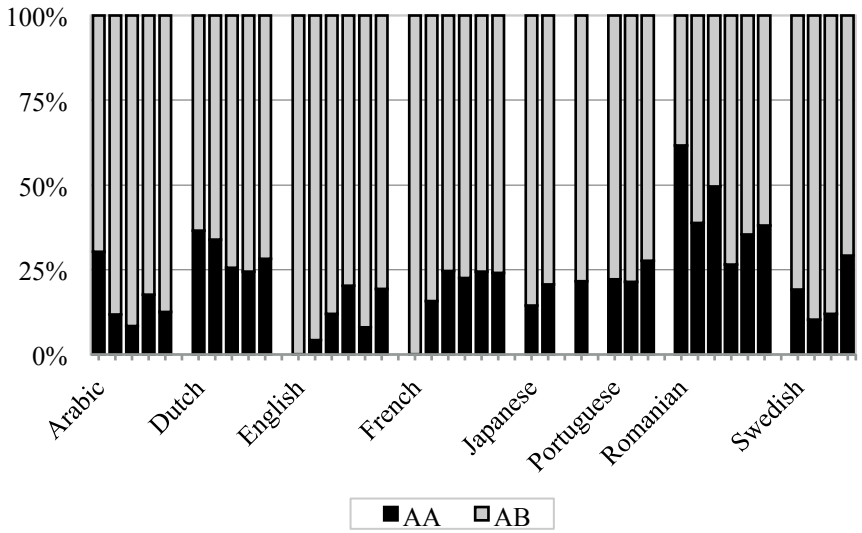


Figure 4.4: Development of two-syllable utterance patterns from seven to 12 months old across all languages. Y-axis label indicates the beginning of the time course for the language.

(χ^2=11.3, df=6, p=0.08), or of month (χ^2=0.8, df=1, p=0.39). However an effect of language was found (χ^2=114.3, df=6, p<0.001). As no interaction was found, the interaction effect was removed from the model, and the same language effect was found. Arabic (Wald=79.37, p<0.001), Dutch (Wald=26.55, p<0.001), French (Wald=12.94, p<0.001), and Swedish infants (Wald=4.63, p=0.03) were found to have significantly higher production rates of this pattern than infants acquiring English.

Finally, for pattern ABC, there was no interaction between month and language (χ^2=3.2, df=6, p=0.78), and no month effect (χ^2=0.7, df=1, p=0.39). However a language effect was found (χ^2= 136.4, df=6, p < 0.001). With the interaction effect removed, the effect of the month and language did not change. Arabic (Wald=14.94, p<0.001) and Romanian (Wald=17.56, p<0.001) differed with respect to this pattern from the baseline English, with Arabic showing significantly more ABC productions than English, and Romanian significantly fewer.

Four-syllable utterances

In comparison with shorter utterances, 4SU are both more numerous and more complex with respect to the categorization of patterns. Therefore, they are not analyzed in the same way as the 2SU and 3SU. Table 4.2 shows how the utterances might be categorized as including both reduplication and variegation. It is not straightforward whether to characterize certain utterances as either reduplicated or variegated as they fall under both categories. For the scope of

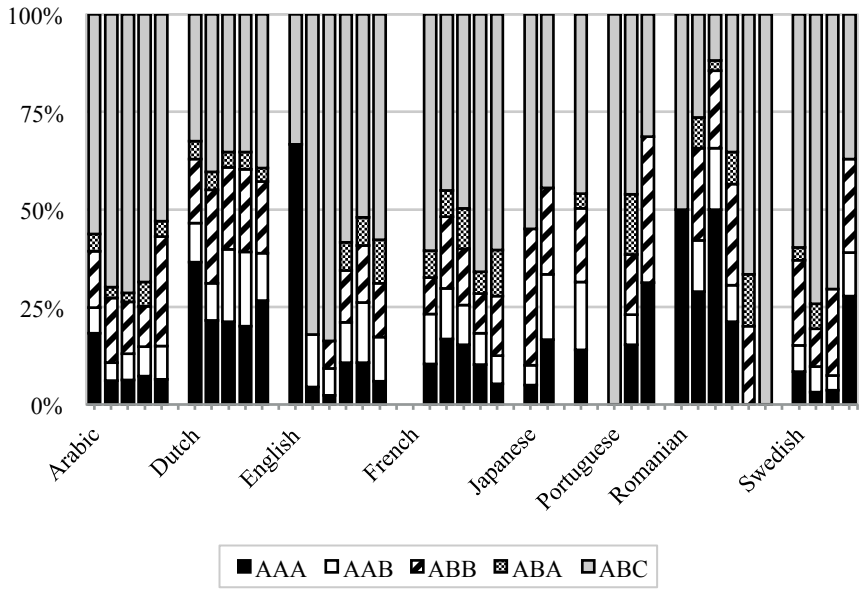


Figure 4.5: Development of three-syllable utterance patterns from seven to 12 months old across all languages. Y-axis label indicates the beginning of the time course for the language.

the present work it thus suffices to present an overview of the results.

Figure 4.6 shows the distribution of all possible 4SU over the eight languages. There are 15 possible patterns. Visual inspection of the figure shows that the pattern with the highest representation across most languages is the completely variegated ABCD. The exceptions are Portuguese and Romanian, which have more instances of the ABBB pattern. In Portuguese, the AAAA pattern is not found at all, while the ABAA pattern occurs 17% of the time. This pattern is found rarely or not at all in all other languages. Dutch also stands out as having the highest frequency of the AAAA pattern, at just under 25% vs. 7.14% in English and 6.33% in French.

In order to get a better idea of the development of 4SU, we grouped the 15 possible patterns into larger categories, depending on whether reduplication was present and where this reduplication occurred in the string. Figure 4.7 shows the grouping of the patterns into six categories, and their frequencies across the represented age groups. When considering any pattern containing reduplication as separate from a full variegation pattern, we can clearly see that full variegation is well represented in most languages. In Arabic, English, French, and Swedish, full variegation thus accounts for more than 50% of the utterances in most age groups (although English, Japanese, and Swedish have fewer variegated utterances over time, and Swedish contains no full variegation at all in month 12). Dutch on the other hand contains approximately 25% variegated utterances, with most of the 4SU productions containing a reduplicated

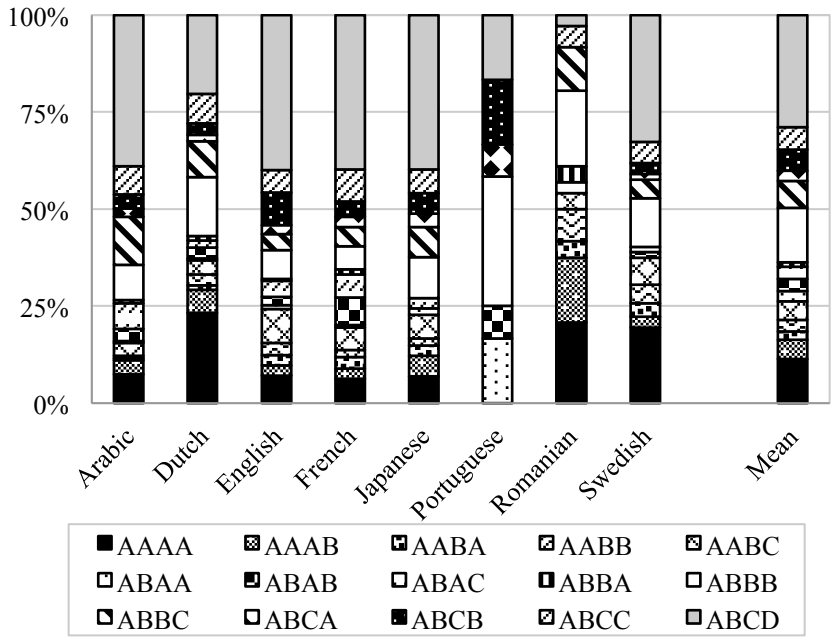


Figure 4.6: Distribution of four-syllable utterances following a totally reduplicated pattern (AAAA), a totally variegated pattern (ABCD) and patterns containing some intermediary level of reduplication.

syllable. Notably Arabic, Dutch, and French do not show a change over time, with their frequency bars remaining constant across age groups.

Thus with the 4SU too, while we did identify some language specific differences, no evidence was found for two distinct stages, nor for a development towards increasingly more variegation.

4.6 General discussion

The present study presents the first attempt at a large-scale and cross-linguistic analysis of infant babbling utterances. Taken together, our results from the 2SU, 3SU, and 4SU show that there is no evidence for clear stages of development across the babbling period with respect to reduplication versus variegation. Month was never an explanatory factor in our models for 2SU and 3SU, while for 4SU, no clear age effects could be discerned either. All age effects interacted with language to show that in certain languages, certain specific patterns (such as AAB) are produced more or less over time. However, in no language did we find the predicted line of development in which reduplicated babbling is most prominent initially and is, either gradually or suddenly, overtaken by variegated babbling. Instead we found that variegated babbling is actually the

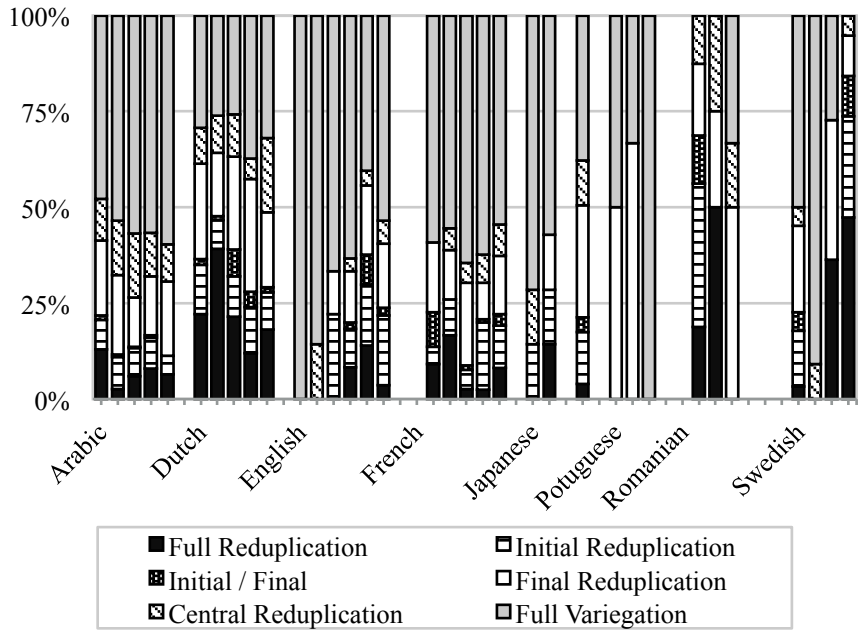


Figure 4.7: Development of four-syllable utterance patterns, grouped by utterance type. Each bar indicates data from one month, from seven to 12 months, across all languages. Y-axis language label indicates location of first age group for that language.

most prominent pattern from the earliest stages of babbling. This was the case in most languages in 2SU, 3SU, and 4SU.

Our data showed language effects on each individual pattern analyzed, showing that there are some important differences between languages with respect to babbling. For example, Dutch differed significantly from English, a closely related language, with respect to the production of completely reduplicated utterances like AA and AAA. Dutch infants had significantly more such utterances than English-acquiring infants. Such between-language differences raise many specific questions related to the connection between each language and the frequency of various patterns. This presents an interesting point of focus for future work. The linguistic profile (the phonology and the morphology) of the native language may have an important role in shaping the types of babbling patterns infants produce and the distribution of these patterns.

Finally, the work presented here may also be continued and expanded by identifying exactly the types of syllables making up reduplicated and variegated utterances. For example, is the variegation or reduplication taking the form of a consonant change, a vowel change, or both? Does this change over time, and does it differ between languages? Both reduplication and variegation may take many forms. In the present work, the patterns were identified on the basis of exact identity, yet this is another area where different language backgrounds

may yield different results.

Our large-scale analysis has thus shown that while the types of patterns infants produce differ depending on their language background, the idea of discrete stages may be abandoned in favor of more fruitful research into explaining particularities of babbling patterns in different language backgrounds.

4.7 Acknowledgements

We are grateful to all researchers contributing to the PhonBank database and to Greg Hedlund and Yvan Rose for their help in scripting and implementation in Phon. We also thank Bingjing Gu and Shenmin Wang for their help with statistical analyses.

CHAPTER 5

Rule learning in adults in the speech domain

This chapter is based on a paper in preparation for submission: Geambaşu, A., Spierings, M.J., ten Cate, C., & Levelt, C.C. (in preparation). Tell me what to learn: The effects of task-specific variables on artificial grammar learning and generalization

The goal of this chapter was twofold. The first goal was to bridge the work conducted with infants in the previous chapters with the work conducted with birds within our larger interdisciplinary project (Spierings, 2016). To facilitate cross-species comparison, we investigated two experimental paradigms, which would respectively be more similar to the ones used in each of these populations. The second goal was to thoroughly investigate the role of various experimental manipulations on rule learning abilities.

5.1 Abstract

Extraction and generalization of rules from a complex input with an underlying structure is one of the bedrocks of language learning. The ability to learn such rules is often studied using the artificial grammar paradigm. In order to understand the conditions under which adults are able to extract these artificial XXY/XYX-type grammar rules and generalize them to novel input, we conducted an artificial grammar-learning task with grammars similar to those used by Marcus et al. (1999). Eight participant groups were exposed via passive familiarization and tested in a Yes/No paradigm. Another three groups were exposed via reinforced training and tested in a Go-left/Go-right task. Between the

groups, we systematically varied the experimental instructions and the number of exposure items and test items. Our results indicate that participants can learn the rule underlying the grammar only under specific familiarization and testing conditions. In this series of experiments, adults do not generalize the underlying rule to novel items, unless they receive indications about what they should attend to, either through explicit instructions, by being presented only with novel test items (directed testing), or with feedback training. Our findings indicate that details of design and instructions in AGL tasks have a great impact on participants' learning and generalization.

5.2 Introduction

Extraction and generalization of rules from a complex input with an underlying structure is one of the bedrocks of language. This ability seems to be present already at birth. Several studies have shown that infants as young as seven months old are able to learn simple XYX, XXY, and XYY rules (henceforth referred to as XYX-type rules[1]) from a very short (two- to three-minute) passive familiarization phase (e.g. Gerken, 2006; Marcus et al., 2007, 1999)[2] and neurophysiological work has shown that even newborns are sensitive to violations in these simple patterns (Gervain et al., 2008). This skill seems to remain present throughout life, as adults are still able to learn and generalize grammars of varying structures and complexity presented both in the auditory and the visual domain. This has been shown using the simple XYX-type rules (e.g. Christiansen et al., 2000; Endress et al., 2007; Sun et al., 2012), non-adjacent dependencies (e.g. Creel, Newport, & Aslin, 2004; Gebhart, Newport, & Aslin, 2009; Gómez, 2002; Kuhn & Dienes, 2005; Newport, Hauser, Spaepen, & Aslin, 2004; Uddén, Ingvar, Hagoort, & Petersson, 2012; Romberg & Saffran, 2013), and complex finite-state and phrase structure grammars (e.g., Fitch & Hauser, 2004, although see Hochmann, Azadpour, & Mehler, 2008; Friederici et al., 2006; Gómez et al., 2000; Petersson, Folia, & Hagoort, 2012; de Vries, Christiansen, & Petersson, 2011; de Vries, Monaghan, Knecht, & Zwitserlood, 2008; Reber, 1967, 1976; Reber et al., 1980. Although abstract rule learning of patterns does seem to be subject to certain perceptual and memory constraints (Endress et al., 2007, 2005), often little attention is given to the experimental factors influencing how well humans perform in such tasks and how they may interact with these perceptual and memory constraints. A lack of standardization across paradigms yields incomparable results and hinders a theoretically cohesive account of rule generalization. The present work uses XYX-type grammars to study the effects of the following experimental factors as sources of potential variation in rule learning ability: the amount of variety in the input, the presence or absence of feedback, the duration of exposure, the instructions given to

[1] In chapter 2, these are referred to as Marcus rules.
[2] Although see chapter 2 of this thesis.

the participant before exposure and/or before test, and the level of activeness or passiveness required of the participant in the task.

Generalization has been shown to be aided by increased variety in the input, and hindered by longer exposure. As Peña, Bonatti, Nespor, and Mehler (2002) showed, increasing the length of exposure to a non-adjacent dependency (AXC rule) by repeating the same sounds did not help to reveal structure in an input stream, but instead strengthened the memory traces for those often-repeated specific items. Fewer repetitions, on the other hand, allowed for rule learning. Thus, it seems to be the variety in the input and not the absolute duration or repetition of exposure that is crucial for rule learning and generalization (see also Endress & Bonatti, 2007, using a version of the AXC grammar, and Johansson, 2009, using a finite state grammar). In the present study, we manipulated the amount of variety in an XYX-type rule learning task, with the expectation that more variety in the input would also lead to better generalization of these types of rules.

Another factor that may play a role in learning is the presence or absence of feedback. The most prevalent types of learning paradigms used in artificial grammar learning experiments are either passive familiarization or training paradigms with feedback. The effect of these paradigm-specific factors on participants' learning strategies and generalization ability has not yet been systematically explored. While a multitude of passive learning tasks have shown that passive rule learning is possible (e.g., Forkstam, Elwér, Ingvar, & Petersson, 2008; Frank, 2013; Peña et al., 2002; Perruchet, Tyler, Galland, & Peereman, 2004; Reber et al., 1980), feedback has been shown to boost learning, as it allows learners to rapidly test, and either reject or accept, hypotheses about the input online. As rule learning has predominantly been considered a task that is resolved by implicit learning, there have been few claims that feedback is necessary for rule learning. Yet feedback has been shown to boost performance in rule learning tasks. Mealor and Dienes (2013) showed that implicit learning of a (visually-presented) finite state grammar is indeed possible both with and without feedback, although only learning with feedback allowed participants to make their knowledge explicit and maintain it longer. Thus, while learning a grammatical rule does seem to be possible without feedback, such feedback may play an important role in the encoding and retention of the knowledge. Learning of the XYX-type rules has almost exclusively been shown using passive familiarization, both with infants (Gerken, 2006; Kovács & Mehler, 2009; Marcus et al., 2007, 1999) and with adults (Christiansen et al., 2000; Sun et al., 2012), although see Endress and Bonatti (2007) for a training-based oddball learning paradigm and Chen et al. (2015) for a training-based categorization paradigm. In the present work, we directly compared two tasks, one with and one without feedback during the learning phase, with the expectation that both conditions would allow for learning of the XYX-type rules, but that feedback would yield significantly better learning.

In addition to feedback, the type of instructions participants receive may affect the strategies they use when faced with a rule learning task. Reber (1976, 1989) showed that finite state grammars are better learned with neutral instructions (telling participants they are taking part in a memory task in which they have to reproduce letter sequences), and argued that such complex grammars are better learned without explicit instructions. However, the influence of instructions on the learning of simple XYX-type rules is not clear, as this has until now not been explicitly investigated. In addition, the fact that there is no uniformity across recent rule learning studies with respect to instructions also makes this important to explore. Within the broader artificial language learning literature, little attention to, or justification for, the chosen form of instruction is given. Across studies, instructions range from vague and uninformative to very explicit. For example, participants are simply told to listen or look carefully in Conway and Christiansen (2006) and Fitch and Hauser (2004); they are told that they are participating in a study on perception and memory in Johansson (2009); they are told they will see nonsense items and are asked to spell them out in Whittlesea and Dorken (1993); they are told that a series of learning stimuli would be presented in Sun et al. (2012). In other experiments, for example, in Gebhart, Newport, and Aslin (2010), participants are explicitly told to find the pattern; they are told they are participating in a pattern recognition experiment and should not pay attention to the individual sounds in Christiansen et al. (2000); in Forkstam et al. (2008) and Gómez et al. (2000), they are told that the language is generated by a complex set of rules; and in Endress and Bonatti (2007), participants are told to answer whether the test items follow the same grammar as in the exposure. In a number of papers, instructions are not explicitly reported at all, or only in vague terms. Of the three examples of XYX-type auditory rule learning experiments with adults, two (Christiansen et al., 2000; Endress & Bonatti, 2007) provide their participants with explicit instructions, while the other (Sun et al., 2012) provides their participants with vague instructions. While justification is not given for the chosen instruction type, piloting results may have been influential in these cases. Notably, Christiansen et al. (2000) used extremely explicit instructions in their replication of Marcus et al. (1999)'s experiment, because piloting showed that their adult participants did not succeed in generalizing the XYX-type rules to novel stimuli with vague instructions (Christiansen, person commun.). While these findings were not the focus of their work, the fact that participants did not readily generalize the simple rules that "even babies" could generalize without instructions or feedback raises the larger question of the conditions under which these types of rules are learnable.

In the present study, we aimed to shed light on how the above-mentioned factors interact and influence adults' ability to learn and generalize simple rules. To date, we find a combination of these experimental factors presented without explicit justification. In our study, we present the first systematically structured design to disentangle the different influences on the learning of sim-

ple XYX-type rules by adults. We manipulated exposure grammar (XXY or XYX), variety in the learning phase (either three or 15 different exposure triads), instructions with different degrees of explicitness, the presence or absence of feedback during the learning phase, and types of test items (either a mix of familiar and novel sounds, or only novel sounds). While we expected adults to be able to generalize these simple rules in every condition, we predicted that more specific instructions would lead to stronger generalization as per Christiansen et al.'s (2002) results (also Christiansen, person commun.) and as per Reber's (1989) conclusion. In addition, we predicted that participants would focus on the surface forms of the individual syllables rather than on the underlying rule when they would be presented with less variation in the familiarization phase, keeping the duration of the familiarization phase constant. On the other hand, we expected that with more variation, participants would not be able to memorize individual syllables or triads, and thus be more prone to generalize the rule. Finally, we predicted that training with feedback would produce the best generalization and learning outcomes.

We tested these hypotheses in a series of four experiments. Experiment 1 is labeled Undirected Familiarization because in this experiment, the instructions given before the familiarization do not give any indication that there is an underlying rule to be found. Experiment 2 is labeled Directed Familiarization because here, participants do receive explicit instructions about what they should pay attention to, directing their attention to the underlying rule. We expected better rule generalization in Experiment 2 than in Experiment 1. For these experiments, the test items consisted of both new combinations of familiar syllables (henceforth combination triads) and combinations of novel syllables not heard before (henceforth novel triads). Experiment 3 is called Implicitly Directed Testing due to the type of test items presented: here participants are only tested with novel syllables instead of with both familiar and novel syllables as in Experiments 1 and 2. We hypothesized that in Experiment 3, participants' attention to the rule would be enhanced if the surface forms in the learning and test phases did not overlap. In one condition, participants received directed instructions before familiarization, while in the other condition they received undirected instructions. Finally, Experiment 4 is called Feedback Training, as in this experiment, participants are trained until criterion with feedback on their response to each learning trial. We hypothesized that generalization to novel items would be the most successful when participants received feedback during training. The manipulations of the four experiments are outlined in Table 5.1. Overall we found that despite the simplicity of the to-be-learned rules, the participants' attention needed to be directed towards the relevant level of processing in order for learning and generalization to occur, either through feedback, explicit instruction, or by cues in the stimuli such as variation in the input or directed testing by only playing novel stimuli.

Table 5.1: Overview of the four experiments presented in the paper, with each line representing one condition. Each condition involved manipulations to the learning paradigm (passive familiarization or training with feedback), testing paradigm (yes/no or go-left/go-right), instructions (referring to language, group, or pattern), types of test items (combination and/or generalization), and number of triads (15 or 3) presented to the participants. Each condition included 16 unique participants.

Experiment		Learning Paradigm	Testing Paradigm	Test Items	Instructions	Triplets
1	Undirected Familiarization	Familiarization	Yes/No	Comb + Gen	Language	15
						3
					Group	15
						3
2	Directed Familiarization				Pattern	15
						3
3	Implicitly Directed Testing			Gen	Language	15
						3
					Pattern	15
						3
4	Feedback Training	Feedback Training	Go-Left / Go-Right	Comb + Gen	-	15
						3

5.3 Experiment 1 - Undirected familiarization

5.3.1 Methods

Participants

Sixty-four adult participants (26 males, 38 females, age M±sd: 25±6.9) were tested in this experiment, equally divided between four conditions. They were all native Dutch speakers and had no dyslexia or hearing problems. Recruitment was conducted via the Leiden University participant recruitment site and the experimenters' personal networks. Participants received either monetary compensation or course credits for participation.

Procedure

The experiment consisted of a familiarization phase followed by a test phase. Participants sat behind a computer screen and wore the provided headphones (Sennheiser HD 201). They were assigned to one of two instruction conditions: Language or Group. In the case of the Language condition, participants received the following instructions: *You will now hear fragments of a non-existing language. Listen carefully. In the test phase you will have to indicate whether the fragment you hear is part of this language.* In the Group condition, participants heard no information about what to listen to before the familiarization: *You will now hear different fragments. Listen carefully.* During the subsequent fa-

miliarization phase, participants heard syllable triads separated by short pauses (1 sec). The familiarization triads followed either an XYX grammar or an XXY grammar. One group of participants was exposed to 15 different triads all repeated three times, another group to three different triads, all repeated 15 times. The familiarization phase lasted approximately two minutes in both exposure conditions.

During the test phase, participants were instructed to score several triads on whether or not they belonged to the same language or group as the triads they were familiarized with. In the Language condition, the instructions were as follows: *Indicate whether the following fragments belong to the same language as in the listening phase. For yes press 'z'. For no press '/'. Respond as quickly and accurately as possible.* Participants in the Group condition were instructed as follows: *Indicate whether the following fragments belong to the same group as in the listening phase. For yes press 'z'. For no press '/'. Respond as quickly and accurately as possible.* During the test, 18 test triads were presented twice, separated into two test blocks, such that in total 36 test trials were heard. Within the two blocks, the test triads were presented in random order. There was no break between blocks. The experiment lasted approximately 10 minutes.

Upon completing the experiment, participants were requested to fill out a questionnaire asking what strategy they used for solving the task (what they paid attention to), as well as their age, sex, languages spoken and whether and to what degree they had been diagnosed with dyslexia or hearing problems.

Stimuli

Familiarization. The stimuli were identical to those used in chapter 2, Experiment 1. They consisted of eight consonant-vowel syllables (*mo, do, ka, sa, pu, nu, le, fi*) produced by a phonologically-trained female native speaker of standard Dutch (C.C.L.). All syllables were equalized in length, average pitch, and amplitude using Praat (Boersma & Weenink, 2014). They were concatenated into triads following two grammars, XYX and XXY, with 100 ms pauses between each syllable (Figure 5.1). Each syllable was assigned to a different letter (A through H) for each participant, to avoid pseudo replication (e.g., syllable A could be *ba* for one participant, but *fi* for another, see chapter 2, Appendix A). Triads formed by syllables A through F were then randomized and concatenated into a continuous stream with 1 sec pauses between each triad. Depending on the experimental condition, either 3 different triads were repeated 15 times or 15 different triads were repeated 3 times in the stream (from now on referred to as 3 triads or 15 triads conditions).

Test

The test stimuli consisted of both triads with an XYX pattern and triads with an XXY pattern. These triads either consisted of new combinations of the syllables used for familiarization (combination triads), or of new syllables (novel

triads; Table 5.2). The two syllables that were not used during familiarization (G and H) were used to create the novel triads and were therefore different for each participant. Test triads were concatenated in the same fashion as the familiarization triads, with 100ms pauses between the syllables.

See Figure 5.2 for a list of all triads used in familiarization and test, and see chapter 2, Appendices A and B for a full list of letter-to-syllable correspondences.

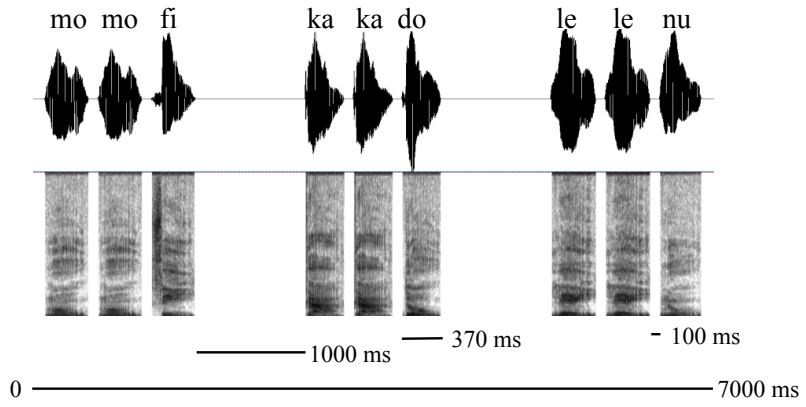


Figure 5.1: Example of three triads of the familiarization stream. Length of pauses between syllables (100 ms), pauses between triads (1000 ms) and syllables (370 ms) are indicated.

Analysis

A percentage correct (%C) score was calculated per participant for their individual responses to the test triads, by adding the correct yes responses to the correct no rejections, and dividing their sum by the total number of trials. This resulted in a score between 0 and 1, with 1 indicating perfect discrimination (all triads were correctly marked as either belonging to the same language/group or not) and 0.5 indicating performance at chance level.

The two types of instructions (Language vs. Group) and the two types of familiarization (3 triads vs. 15 triads) created four experimental conditions. We analyzed the effect of these variables with an LMM. The dependent variables were the %C of the novel triads and the combination triads. The covariates were the type of familiarization grammar (XYX or XXY) and the sex of participants, with subject nested in experiment as random factor. Interactions between the independent variables were measured with a linear regression afterwards. Furthermore, per experimental condition, we tested whether the %C scores deviated from random (0.5) with a one-sided t-test.

The reported strategies of the participants were divided into three categories: pattern, sounds, or other. This division was based on the usage of these

Table 5.2: Triads used for familiarization and the Combination and Generalization test conditions. Letters represent a consonant-vowel syllable, which was different for each participant. Bold familiarization triads were used in the 3 triads familiarization condition. Combination triads were not used in Experiment 3. See chapter 2, Appendices A and B for letter-to-syllable correspondences and full strings.

Familiarization triplets		**Combination triplets**		**Generalization triplets**	
XYX	**XXY**	**XYX**	**XXY**	**XYX**	**XXY**
ADA	AAD	ACA	AAC	GHG	GGH
AEA	AAE	BDB	BBD	HGH	HHG
AFA	AAF	CEC	CCE		
BAB	BBA	DCD	DDC		
BCB	BBC	EAE	EEA		
BFB	BBF	FAF	FFA		
CAC	**CCA**	FEF	FFE		
CDC	CCD				
CFC	CCF				
DBD	DDB				
DED	DDE				
DFD	**DDF**				
EBE	**EEB**				
ECE	EEC				
EFE	EEF				

particular words or words with similar meaning by the participants. Reported strategies categorized as 'pattern' included reports with any of the following words: pattern, rhythm, order, and combination. The strategies grouped into the category 'sounds' involved self-reported listening for any of the following: sounds, words, vowels, and consonants. The category 'other' included self-reported strategies such as: feeling and intuition. These reported strategies were tested in an ANOVA for an effect on the %C score for the novel triads. Post-hoc t-tests were used to test which strategies were related to differences in %C scores. The same types of analyses were conducted for all four experiments presented in this paper. The variables used in the LMM are explained separately for each experiment.

5.3.2 Results

Combination triads

The model showed no effect of instruction (Group: M = .66, Language: M = .71, F = 1.39, p = 0.24) or the number of familiarization items on the %C score for the combination triads (3 triads: M = .64, 15 triads: M = .73, F = 2.61, p = .12), nor an interaction effect of these two factors (F = .12, p = .73, Figure 5.2a). Participants' sex or the familiarization grammar (XXY or XYX) also did not explain the output variance (Sex: F = 0.05, p = .82; Fam. grammar: F = 2.68, p = .11). Variable reduced models increased the statistical power but did not alter the results of this analysis.

In all four experimental conditions, subjects correctly discriminated the grammars consisting of combination triads with known syllables (Language 3 triads: M±sd = .68±.2, p = .002; Language 15 triads: M±sd = .74±.2, p < .001; Group 3 triads: M±sd = .59±.17, p = .004; Group 15 triads: M±sd = .72±.2, p = .001; Figure 5.2a).

Novel triads

Also for the novel triads, there was no effect of instructions or number of familiarization items on the discrimination performance (Instructions: F = 0.313, p = .58; No. of triads: F = 0.08, p = .78).

One-sided t-tests showed that in each experimental condition subjects were, on average, unable to discriminate correctly between the grammars, as the results did not differ from performance at random (Language 3 triads: M±sd = .62±.2, p = .2; Language 15 triads: M±sd = .55±.3, p = .48 Group 3 triads: M±sd = .59±.2, p = .07; Group 15 triads: M±sd = .6±.2, p = .11; Figure 5.2b).

5.3.3 Reported strategies

Reported strategies of all participants showed two experimental conditions in which the participants mostly paid attention to the structure of the triads (Structure strategy). These were Group instructions with 3 triads and Language instructions with 15 triads. In the other two conditions participants reported slightly more Sound strategies, or both strategies were reported equally often (Figure 5.3). Overall there was no effect of instructions or number of triads on the reported strategies (Instructions: F = 1.69, p = .43; No. of triads: F = .78, p = .38). Reported strategy was correlated with the %C score for the generalization test items. Participants who reported using a Structure strategy had a higher %C compared to participants who reported using a Sound strategy (Structure strategy: M±sd = .68±.24; Sound strategy: M±sd = .56±.15, F = 3.2, p = .046).

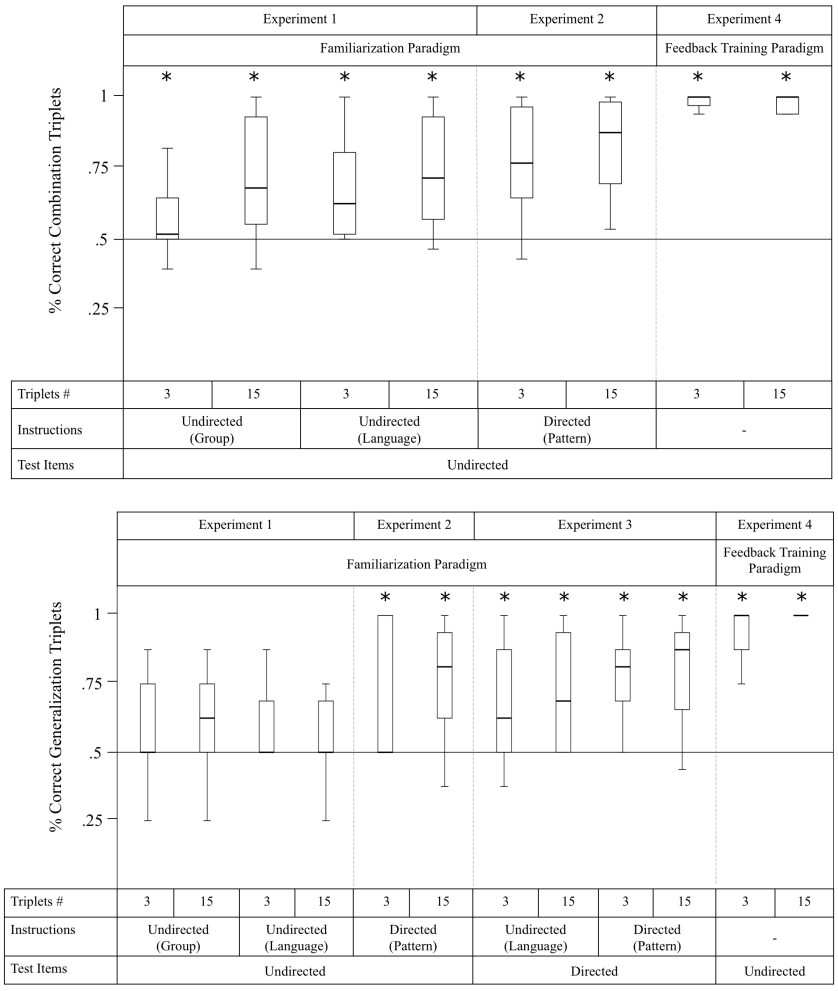


Figure 5.2: a) %C scores to the Combination test triads for each experimental condition. b) %C scores to the Generalization test triads for each experimental condition. In both 2a and 2b, whiskers reach to the minimum and maximum score.

5.3.4 Discussion

When instructed to indicate whether triads belong to the same Group or Language as previously heard triads, participants did not generalize the structure of triads with novel syllables. The strategy that they reported to use was independent of the amount of variation presented during familiarization. Neverthe-

less, participants who reported to have focused on structure had a significantly higher %C than those who focused on sounds. This suggests that focusing on the segmental make-up of the syllables hampers the acquisition of the underlying rule.

This finding leads to the hypothesis, tested in the next experiment, that directing participants' attention towards the pattern in the familiarization triads will facilitate rule learning. Instructing participants before familiarization to pay attention to the pattern in the stimuli should therefore enhance overall rule generalization scores.

5.4 Experiment 2 - Directed familiarization

5.4.1 Methods

Thirty-two adult participants were tested, equally divided over the Combination and Generalization conditions (male: 12, female: 20, age M±sd = 23.7±6.6). They were all native Dutch speakers and had no dyslexia or hearing problems. Participant recruitment and compensation was the same as in Experiment 1.

Procedure

Except for the instructions to the participants, the procedure, stimuli, and analysis of this experiment were all identical to Experiment 1. Instructions before familiarization read: *You will now hear fragments that follow a specific pattern. Listen carefully. In the test phase you will have to indicate whether the fragment you hear follows the same pattern.* Before the test, instructions read: *Indicate whether the next fragments follow the same pattern as the fragments in the listening phase. For yes press 'z'. For no press '/'. Respond as quickly and accurately as possible.*

5.4.2 Results

Combination triads

The LMM showed that the variation between the two conditions was not due to variation in the number of familiarization triads ($F = .74$, $p = .40$). There was no effect of participants' sex or the familiarization grammar (Sex: $F = 1.1$, $p = .31$; Fam. grammar: $F = 1.02$, $p = .32$). In both the 3 triads and 15 triads conditions, participants were able to correctly discriminate between test triads following the familiarization pattern and those following the alternative pattern (3 triads: M±sd = .79±.19, $p < .001$; 15 triads: M±sd = .83±.18, $p < .001$; Figure 5.2a).

Novel triads

Both groups correctly generalized the pattern to novel triads (3 triads: M±sd = .68±.24, p = .006; 15 triads: M±sd = .77±21, p < .001; Figure 5.2b). No difference between the %C to novel triads was found between the two conditions, indicating that there was no effect of number of familiarization triads (F = 1.69, p = 0.2). For these items there was no effect of participants' sex or the familiarization grammar (Sex: F = 3.11, p = .09; Fam. grammar: F = 1.78, p = .19).

Reported strategies

The reported strategies showed that participants had a clear tendency to pay attention to the structure instead of the specific sounds, in both experimental conditions (Figure 5.3). There was no difference in reported strategy between the two conditions, nor were the strategies dependent on age or sex of the participant (Age: F = 1.3, p = .34; Sex: F = 1.74, p = .21). The %C score to novel triads was strongly correlated to the reported strategy. Participants who reported using the Structure strategy had a higher %C score in both experimental conditions (Structure: M±sd = .80±.24; Sound: M±sd = .51±.13, F = 5.7, p = .008).

5.4.3 Discussion

Under the new instruction conditions of this experiment, adults were able to generalize the rule underlying the stimuli. However, a number of participants judged the novel test stimuli as having a different pattern. Notably, these were the participants who reported to have used the Sound strategy. While giving directed instructions did lead to an increase in the number of participants that learned the underlying rule, it was still not a strong enough condition for all participants to learn the rule.

Up until now, during the test phase, participants were presented with a mixture of triads, formed either from familiar syllables or from novel syllables. When in the test phase triads with familiar syllables happen to be presented first, a participant might be driven to adjust the Structure strategy – induced by the instruction – to a more limited strategy, namely, that only familiar patterns with familiar syllables should be judged as similar to the familiarization triads. If the test items would not contain familiar syllables at all, it might be easier for participants to focus on the structure of the test triads. In the next experiment, we examine this hypothesis further.

5.5 Experiment 3 - Implicitly directed testing

5.5.1 Methods

Participants

Sixty-four adults participated in this experiment, equally divided over four experimental conditions (males: 25, females: 39, Age M±sd: 22.6±5.4). They were all native Dutch speakers and had no dyslexia or hearing problems. Participant recruitment and compensation were identical to Experiments 1 and 2.

Procedure

The general procedure was the same as in the previous experiments. Participants were familiarized with the same triads as in Experiments 1 and 2. Half of the participants heard the Language instructions from Experiment 1, while the other half heard the Pattern instructions from Experiment 2. As in the previous experiments, the familiarization stream consisted of either 3 triads repeated 15 times, or 15 triads repeated 3 times. In contrast to the previous experiments, in the test phase participants were now only presented with novel triads. The test phase consisted of four blocks, each consisting of four different novel triads (GHG, HGH, GGH, HHG) presented in random order.

5.5.2 Results

Novel triads

Pattern instructions elicited a significantly higher %C score compared to Language instructions ($F = 4.1$, $p = .04$). No effect of number of familiarization triads ($F = .06$, $p = .8$) and no interaction was found ($F = .01$, $p = .93$). No effect of the participants' sex or the familiarization grammar was found (Sex: $F = 3.29$, $p = .07$; Fam. grammar: $F = 0.03$, $p = .87$, Figure 5.2b).

In all four experimental conditions, participants discriminated between triads following the consistent and those following the inconsistent grammar (Language 3 triads: M±sd = .69±.22, $p = .004$; Language 15 triads: M±sd = .71±.2, $p = .001$; Pattern 3 triads: M±sd = .78±.15, $p > .001$; Pattern 15 triads: M±sd = .81±.18, $p > .001$, Figure 5.2b).

Test blocks

Participants received four blocks of four randomized test items consecutively, resulting in 16 total test trials. Analysis showed no difference in their %C score between these four blocks (Block 1 M±sd = .77±.24; Block 2 M±sd = .79±.25; Block 3 M±sd = .79±.24; Block 4 M±sd = .75±.20; $F = 1.46$, $p = .24$) indicating that no learning effect throughout the test phase could be observed.

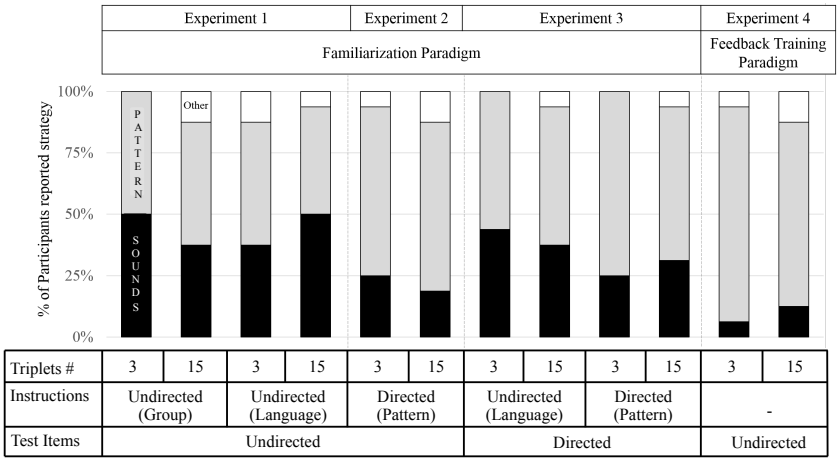


Figure 5.3: Percentage of participants who reported to have used a particular strategy, divided by experiment and experimental condition. Every bar entails 16 participants. Black areas indicate the percentage of participants who reported using a Sound strategy and the grey areas indicate the Pattern strategy users. White areas represent the reports of Other strategies or people who did not report a strategy.

Reported strategies

Reported strategies showed that in all experimental conditions there were more participants applying a Structure strategy than a Sound strategy (Figure 5.3). However, in this case, there was an effect of instruction type; more participants reported to use a Structure strategy in the Pattern instruction conditions than in the Language instruction condition ($p = 0.01$). Again, there was a strong correlation between %C score of the test items and the reported strategy. Participants who reported using the Structure strategy had a higher %C score compared to participants who reported using a Sound strategy (Pattern: M±sd = .84±.17; Sound: M±sd = .61±.15, $F = 16.8$, $p < .001$).

5.5.3 Discussion

Testing with only new items increased the number of participants who correctly categorized these test items. This could be due to the fact that in this experiment participants' attention was directed towards the underlying rule in two different ways, both in the familiarization and in the test phase. Together with the results of the previous experiments, this shows that adults require directed attention in order to be able to generalize rules reliably from familiarization stimuli.

By directing participants' attention towards the rule in both the familiarization phase (through instruction) and in the test phase (by directed testing),

we may have exhausted the possibilities to facilitate adults in their rule generalization abilities in a familiarization paradigm. The next and final step in facilitating rule learning in adults is to provide them with feedback while they are learning. We expect that feedback on each training response allows participants to test and update their hypotheses online. This will boost rule learning, resulting in even better test results than those obtained in Experiments 1 through 3.

5.6 Experiment 4 - Feedback training

5.6.1 Methods

Participants

Thirty-two adults participated in this experiment, equally divided over two experimental conditions (males: 17, females: 15, age M±sd = 25.8±4). They were all native Dutch speakers and had no dyslexia or no hearing problems. Recruitment and compensation was identical to Experiments 1 through 3.

Procedure

Participants sat behind a computer screen on which they saw a symbol of a speaker. They were instructed to click on the speaker symbol in order to start a trial. A syllable triad would be presented, which followed either the XYX or the XXY grammar. After the stimulus was played, a green and a red square appeared on the left and right sides of the screen respectively, without further instructions. Clicking one of these squares resulted in immediate feedback with either a happy smiley face with a ping sound (positive feedback) for a triad following one grammar, or a sad smiley face accompanied by a low rumbling sound (negative feedback) for a triad following the other grammar. Before the test phase, participants had to learn to click on one color after hearing a triad following one grammar rule, and to click on the other color when they heard a triad following the other rule. Which color belonged to which grammar rule was balanced between participants. Training stopped when participants responded correctly in at least 75% of the trials, after which the test started. Participants were instructed that they would not receive feedback anymore. As in Experiments 1 and 2, they now heard both Combination and Novel test triads and continued to indicate whether the triads fell into the green or red category.

The duration of the experiment depended on how many training trials participants needed in order to reach the criterion. The mean duration of the experiment was 10 minutes.

5.6.2 Results

Combination triads

In both experimental conditions, participants had a %C score higher than 0.5 (3 triads: M±sd = .97±.08, $p < .001$; 15 triads: M±sd = .96±.07, $p < .001$, Figure 5.2a). There was little variation within the group, with most participants performing at the highest level (29/32 with a %C > 0.87). The small amount of variation found was not explained by the number of triads, sex, or age (No. of triads: $F = 0.3$, $p = .61$; Sex: $F = 1.94$, $p = .21$; Age: $F = 0.43$, $p = .91$).

Novel triads

The %C score of the novel triads was higher than 0.5 in both conditions (3 triads: M±sd = .91±.17, $p < .001$; 15 triads: M±sd = .98±.04, $p < .001$, Figure 5.2b). Again, the limited variation was not explained by number of triads, sex, or age (No. of triads: $F = 0.46$, $p = .58$; Sex: $F = .91$, $p = .43$; Age: $F = 0.8$, $p = .74$).

Reported strategies

In these experiments, participants showed an overall tendency to report a Pattern strategy with only a few exceptions in each condition (Figure 3). The exceptions were participants who reported to have checked whether a sound 'felt good' or whether it was 'happy or sad.' These participants succeeded in the training but did not generalize the rule to novel test items.

5.6.3 Discussion

Discrimination between triads from different grammars, consisting of either familiar or novel syllables was near the maximum level, meaning that adults readily learned and generalized the rules when they were trained in a feedback paradigm. Performance on this task significantly exceeded performance in Experiments 1 through 3, even though no explicit instructions and no directed testing formed part of the procedure. In addition to the feedback provided in Experiment 4, attention to the rule might have been enhanced by the two types of triads in the training phase, which consisted of the same syllables, arranged either in the XYX or in the XXY pattern. This allowed the participants to rapidly determine that the triads come in two patterns. In short, given the right experimental conditions, adults can rapidly generalize rules from input data.

5.7 Overall analysis

In addition to analyzing the results of each experiment separately, we ran an overall GLM model on the results of all the experiments together. This analysis

gives us a more general view on the effects of the number of triads during familiarization and allows us to compare the training length and the reported strategies between the different experiments.

5.7.1 Stimulus variability

The overall analysis showed that, on average, participants who heard 15 different triads during familiarization had a higher %C score than those who heard only 3 different triads, when tested with combination triads (paired t-test: $t = -6.41$, $p < .001$). This effect was not found in the responses to the novel triads ($t = -0.92$, $p = .36$). However, for the novel triads, more participants in the 15 triads condition than in the 3 triads condition had a %C score above 0.5, (3 triads M±sd = 51±20, 15 triads M±sd = 68±22, $p = .04$). No effect was found of participants' age or sex, nor of the grammar that they heard during familiarization. These results indicate that overall, an increase in the variety of triads during familiarization boosts learning and has a positive effect on generalization.

Table 5.3: Table 3. Overview of the four experiments, with each row representing one condition. In the right-most columns are the % of individuals who performed above chance (above .5 %C) per experiment, per test condition.

Experiment		Learning Paradigm	Testing Paradigm	Test Items	Instructions	Triplets	% Success Combination	% Success Generalization
1	Undirected Familiarization	Familiarization	Yes/No	Comb + Gen	Group	3	50	37.5
						15	81.25	62.5
					Language	3	75	37.5
						15	81.25	43.75
2	Directed Familiarization				Pattern	3	87.5	43.75
						15	100	87.5
3	Implicitly Directed Testing			Gen	Language	3	-	37.5
						15	-	50
					Pattern	3	-	62.5
						15	-	62.5
4	Feedback Training	Feedback Training	Go-Left/ Go-Right	Comb + Gen	-	3	100	87.5
						15	100	100

5.7.2 Training length

Reinforced training (Experiment 4) lasted as long as a participant needed to reach the discrimination criterion. On average, participants needed to hear each positively-reinforced triad 10.6 times in the 3 triads condition (32 total positive trials) to reach training criterion. In the 15 triads condition, the participants only needed to hear each positively-reinforced triad 1.5 times (22.5 total positive trials) to reach criterion. This means that participants in Experiment 4 needed

fewer trials when there were more different triads presented. Compared to the participants in Experiments 1 through 3, who were always presented with either 15 x 3 or 3 x 15 different triads, these participants also required fewer instances of the triads to learn the structure (22.5 versus 45 total familiarization triads, $t = -8.74$, $p < .001$).

5.7.3 Reported strategies

Results show that participants who reported to focus on the pattern of the familiarization or training triads, had a higher %C score in the test than participants listening to the sounds ($F = 19.2$, $p < .001$, Figure 5.4). The reported strategy was partially dependent on the given instructions. When given the Pattern instructions, participants reported using a Pattern strategy more than they used a Sound strategy (Structure = 39; Sound = 22). This is not surprising, as this was precisely what they were instructed to do. The other two instruction types, Group and Language, did not show a difference in number of participants who reported using the Pattern and the Sound strategy (Figure 5.5).

When comparing participants who were able to generalize the rule to new instances, with participants who were not, we found no effect of their sex or age (Sex: $t = -1.64$, $p = 0.11$; Age: $t = 1.32$, $p = 0.14$). Furthermore, the number of languages that participants spoke or their overall language proficiency did not explain the variation within each experiment nor the variation between experiments (No. of languages: $t = 0.89$, $p = 0.74$; Language proficiency: $t = 1.28$, $p = 0.23$).

5.8 General Discussion

In our experiments, we found that adults can correctly identify test stimuli as conforming to the familiarized XYX or XXY pattern when confronted with previously-heard syllables, in all conditions. However, the ability to generalize to novel stimuli was only evident in conditions in which participants' attention had been specifically directed towards the underlying structure during familiarization, either with feedback or directed familiarization (explicit instructions), or in the directed testing condition, in which only novel items were presented in the test phase. At the group level, feedback training elicited stronger rule generalization than passive familiarization. In passive familiarization conditions, rule generalization was strengthened by more specific instructions, or by directed testing.

Our hypothesis that more variation in the input would improve learning was supported: more variation during familiarization led to better performance on combination test items. However, it did not result in significantly better results on novel test items, i.e., on rule generalization. Taken together, these

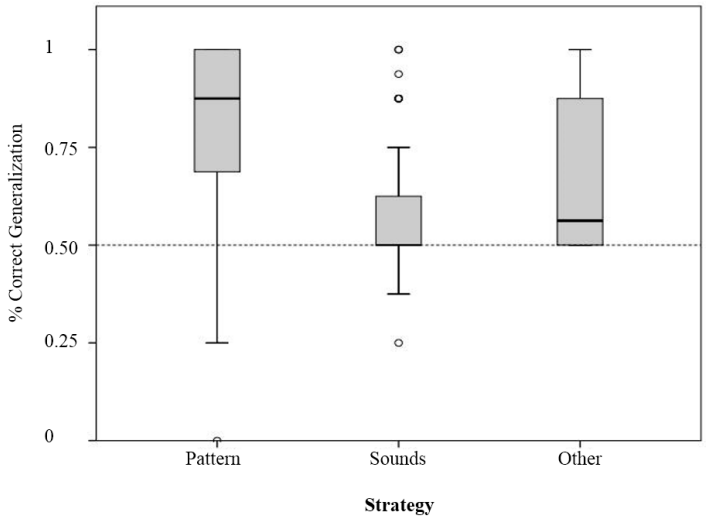


Figure 5.4: %C score to the generalization test items split by reported strategy. Participants who reported to have used a Pattern strategy were on average significantly better rule generalizers compared to participants who reported the use of a Sound strategy.

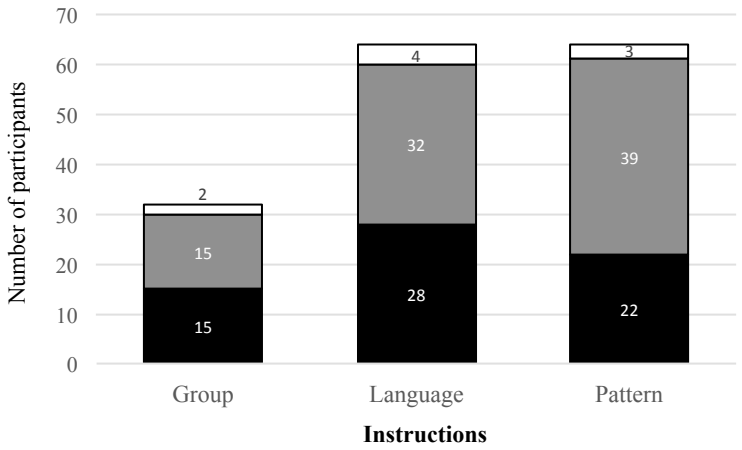


Figure 5.5: Strategies reported by participants when they were instructed with either the Group, Language or Pattern instructions. The black areas indicate the number of participants reporting a Sound strategy, grey areas show the participants reporting Structure strategies, the white areas are the Other strategy participants. Structure strategies involved participants who reported to pay attention to the structure, pattern, or rhythm of the sounds. Sound strategies included participants who reported to pay attention to the sounds, syllables, or words that they heard.

results reveal that adult participants have an initial item-based level of processing of the familiarization stream, that only reaches a more abstract level under guidance. Directing the attention by instructions or feedback helps to limit the numerous strategies that could be employed when processing the familiarization stream. Below, we attempt to explain each of our results in more depth.

5.8.1 Variation

Variation in familiarization items has been shown to be beneficial to rule generalization in at least one type of grammar learning paradigm; non-adjacent dependencies (like in an AXC grammar) are learnable, by both adults and infants, only when the middle item (X) is highly variable (Gómez, 2002). With more variability in the input, the common feature in the data stands out, highlighting the relevant information (see also Onnis, Christiansen, Chater, & Gómez, 2003). However, the role of stimulus variability in the learning of XXY/XYX-type grammars has, up until now, not been investigated. In our experiments, we found that the role of input variability aided in the speed with which participants learned. Participants in Experiment 4 needed fewer trials to reach criterion when trained with 15 triads instead of three. Furthermore, across all experiments, more participants had a score above 0.5 when familiarized with 15 triads than when they were familiarized with 3 triads. Participants who were exposed to more variation also performed better on the combination triads in the test, suggesting that variation in the input speeds up rule learning. Nevertheless, this effect was less clear for the ability to generalize the rule to novel syllables. A possible explanation is related to the specific structure of the familiarization triads. Both in the 15 triad and in the 3 triad condition, six different syllables were used (see Table 5.2). The two conditions thus differed in the amount of variation in the triads (15 versus 3), while the variation in syllables was constant between the two conditions (6 different syllables). This could explain why performance did not increase for test triads formed with novel syllables.

5.8.2 Feedback

Learning without feedback is possible, as was shown in our Experiments 1 through 3 in all Combination tests in all conditions, and in the Generalization test in the Directed Instruction or Directed Test conditions. Yet, the presence of feedback turned out to be the best predictor of success in our study. Participants who were trained with feedback during the familiarization phase (Experiment 4), learned from fewer trials and performed better on both the combination and the generalization test items than participants in experiments where feedback was not provided (Experiments 1-3). The claim that feedback is helpful is widely supported (e.g., Ashby, Maddox, & Bohil, 2002; Dale & Christiansen, 2004; Edmunds, Milton, & Wills, 2015; Maddox, Ashby, & Bohil, 2003; Opitz,

Ferdinand, & Mecklinger, 2011). Specifically, it has been argued that feedback helps to make knowledge explicit and allows participants to retain that knowledge longer (Mealor & Dienes, 2013). The role of feedback is thus to guide participants to a more abstract level of processing. In our experiments, participants who reported to have used a Pattern strategy in the exposure phase had higher %C scores on the generalization triads than participants who reported to have used a Sound strategy. Most of these participants took part in either the Directed Instructions experiment or in the Feedback Training experiment. This is in line with our hypothesis that participants' attention needs to be guided to the relevant features of the task and grammar in some way, either through feedback, directed instructions, or implicitly directed testing.

5.8.3 Instruction

Our results showed that participants were better able to generalize the rule to novel stimuli if they were instructed about the fact that familiarization stimuli followed a certain pattern. At first glance, it may not be surprising that participants provided with directed instructions were able to generalize better than our undirected-instruction participants. Yet this finding is not trivial. In the AGL literature, justification of the type of instructions participants received is often missing if it is not the object of study. While Reber (1976) found that more explicit instructions yield poorer performance in finite state grammar learning, our experiments show the opposite to be true with simpler XYX-type grammars. The role of complexity in implicit or explicit learning was already pointed out by Reber (Reber, 1976, 1989) who stated that explicit instructions about an underlying rule might actually be a hindrance to generalization when the rule is too complex to articulate. This claim has been corroborated by more recent work showing that participants generalized a rule to novel input significantly more successfully for simple, but not for complex, grammars, when instructed to find the rule instead of being told that they were participating in a memorization task (Johnstone & Shanks, 2001; Kuhn & Dienes, 2006). From our experiments, we can also conclude that for the majority of our participants, generalization did not occur without specific instruction. Thus, XYX-type rules were too simple to learn implicitly, yet without directed instruction, explicit learning could not take place either.

Our findings corroborate previous work on the important role of both feedback and instruction in learning. We find that the most crucial factor in the ability to generalize a rule is the direction of attention in the training phase. Failure or success on an XXY/XYX-type rule learning task without feedback or direction in adults seems to depend on the type of information individual participants decide to focus on, leading to success only if this happens to be structural information. This in turn suggests that results obtained with other research populations, such as infants or non-human animals, might be explained by differences in interests and motivations, or by the relative success of the way the experiment is set up to guide attention in the right direction, rather than by

differences in learning abilities. This opens the door for different interpretations of studies looking at rule learning abilities across ages and across species.

5.9 Conclusion

We presented four experiments, with 12 tasks in total. Together they show that rule generalization at the group level is not possible without directing the attention of the participant to relevant features of the stimuli or task in some way, either during the training phase or in the test. The presence of feedback or explicit instructions during training, or of directed testing, turned out to provide the required guidance of attention. The role of instruction and feedback in learning of more complex rules has been explored extensively in the literature, but it has often been ignored in implementations of more simple AGL tasks. In addition, the role of what we have called 'directed testing' has also been neglected in the literature. Our results indicate that stimuli presented in the testing phase continue to play a role in the learning process and in the learners' interpretation of the the task. This finding indicates that participants continue to update their strategy with each trial and modify the mental representation they may have formed during the learning phase accordingly. Our research opens the door for several directions for future research. First and foremost, future research should be aware of (and strictly control for) the task-related factors that have been shown here to influence the outcome of AGL learning experiments. Future work should also further explore the participant-related factors and individual differences in learning. Across all of our experimental conditions, we found individuals who were able to generalize, even in conditions without explicit instruction, directed testing, or feedback. This indicates that rule generalization is possible in individuals without directed attention provided by the task. Across all experiments there were individuals who indicated having attended to the separate syllables of the triads (a Sound strategy) but who were nevertheless able to generalize the rule. This shows that rule learning can occur without explicit knowledge. Reconciling the group tendency and the individual differences with theories of attention, implicit learning, and AGL should be the focus of future work.

5.10 Acknowledgements

We would like to thank Jelle Zuidema and Raquel G. Alhama for valuable feedback during the design and running of these experiments, and for comments on an earlier draft of this paper. Thank you to Renske Kuijpers for suggestions on statistical analyses, and to the students who assisted in data collection: Renske Jacobs, Rabia Mahboeb, Astrid Gilein, Isabella Jordanoska, Gijsbert Westland, Arne Dits, and Sophie Verstraeten.

CHAPTER 6

Rule learning in adults in the auditory domain using Lindenmayer grammars

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

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

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

6.1 Abstract

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

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

6.2 Introduction

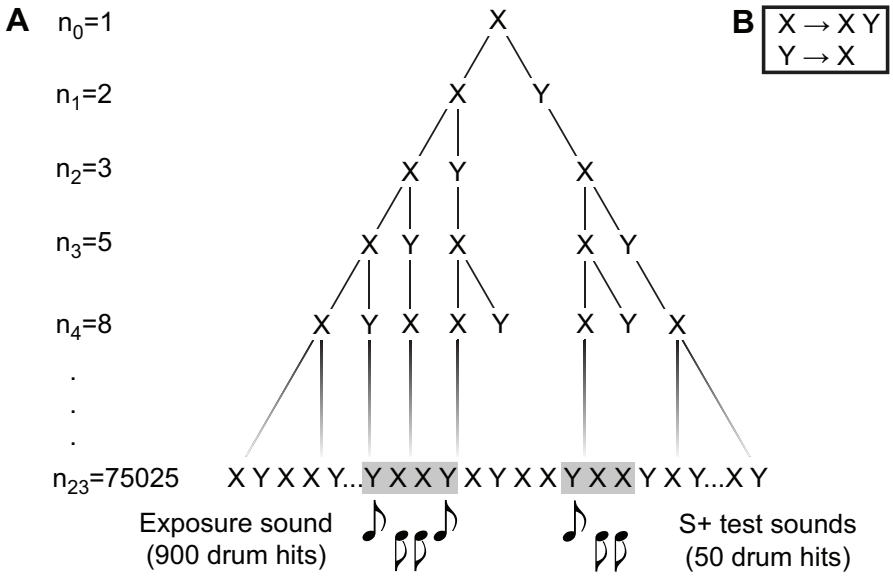
Processing of hierarchical structures has been proposed as a uniquely human ability, a hallmark of the linguistic system that distinguishes human language from animal communication systems (Hauser, Chomsky, & Fitch, 2002; Martins, 2012). Recursion is often considered the pinnacle of human-specific hierarchical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

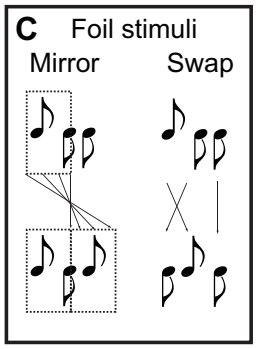
Yet to what extent learning of an A^nB^n grammar can be taken as evidence for processing recursive information at all is debated. Some researchers argue that human participants could in fact use simpler strategies, such as counting and matching the number of As and Bs in a test sequence (Hochmann et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

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

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

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





D

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

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

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

6.3 Experiments 1 and 2

6.3.1 Methods and materials

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

Participants

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

Stimuli

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

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

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

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

Materials

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

Procedure

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

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

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

6.3.2 Descriptive statistics and results

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

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

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

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

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

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

6.3.3 Discussion

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

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

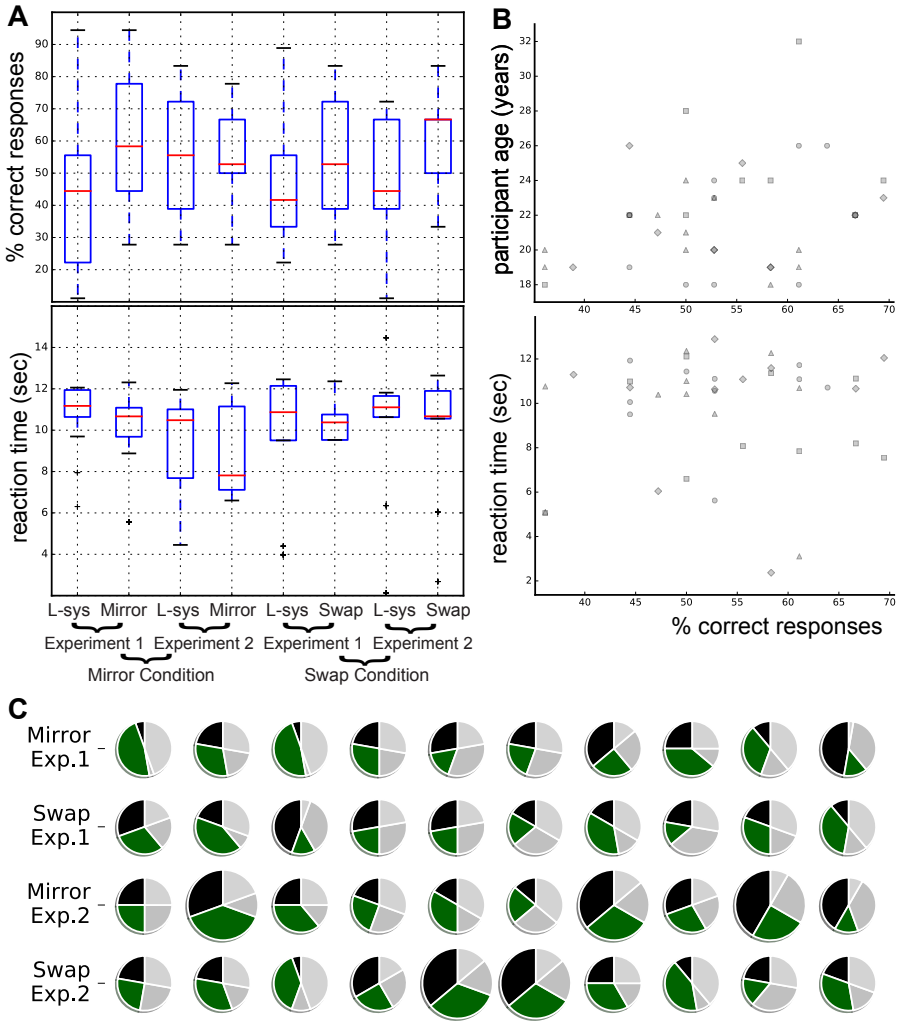


Figure 6.2: Summary of participants' performance at group (A) and individual (B,C) level. (A) Boxplot of percentage correct responses by experimental condition (Mirror vs. Swap), experiment number (limited vs. detailed instructions), and stimulus type (L-sys denotes a correct acceptance of a grammatical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

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

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

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

[1]See also chapter 5 of this thesis.

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

6.4 Experiment 3

6.4.1 Methods and materials

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

6.4.2 Participants

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

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

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

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

6.4.3 Stimuli

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

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

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

6.4.4 Materials

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

6.4.5 Procedure

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

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

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

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

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

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

6.4.6 Analysis of the data

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

6.4.7 Results

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

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

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

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

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

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

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

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

Binomial test

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

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

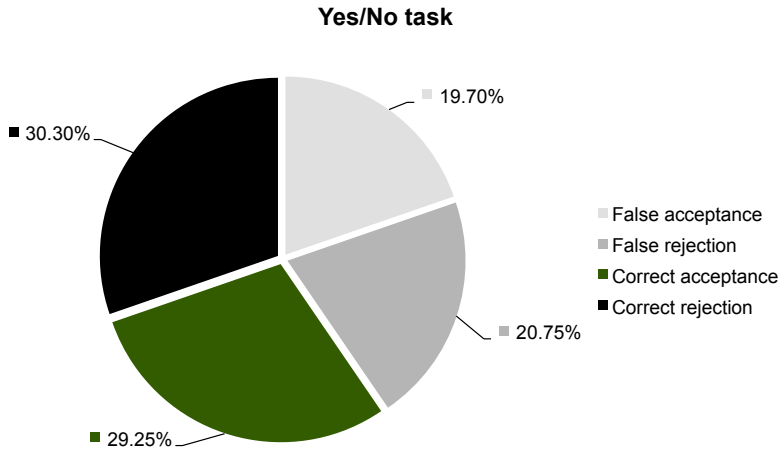


Figure 6.3: Correctness by Category in the Yes/No task. Correct acceptance and correct rejection were both respectively more common than false acceptance or false rejection.

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

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

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

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

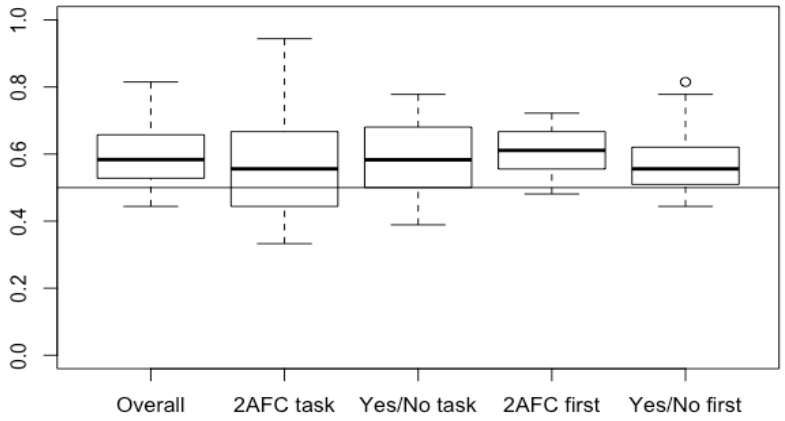


Figure 6.4: Proportion correct of participants of all data, split across tasks, and split across conditions (line shows 0.5 chance level).

Logistic regression

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

6.4.8 Discussion

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

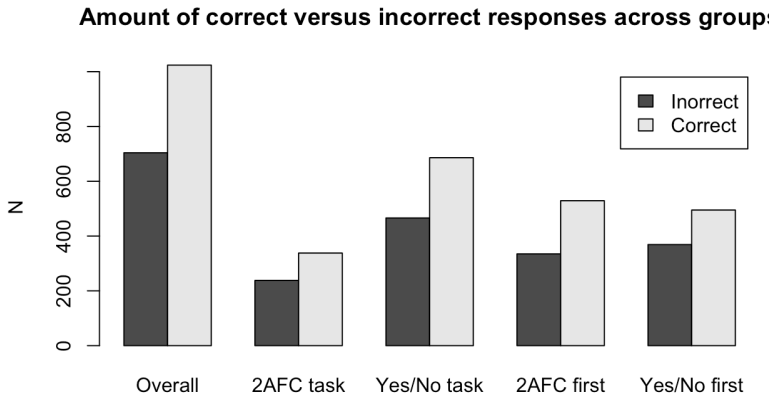


Figure 6.5: Number of correct versus incorrect responses per task and per condition. In all cases, participants provided more correct responses than incorrect responses.

6.5 General Discussion

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

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

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

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

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

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

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

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

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

6.6 Acknowledgments

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

6.7 Appendices

6.7.1 Appendix A

Overview of the data files and their formats

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

Variable names and coding (values in brackets):

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

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

CHAPTER 7

Conclusion

7.1 Thesis summary

The goal of this thesis was to gain new insights into the rule learning capabilities of humans, both infants and adults.

In chapter 2, we studied the ability of infants between the ages of six and nine months to learn grammars similar to the ones used in Marcus et al. (1999). In Experiments 1 and 2 of that chapter, where we used natural speech stimuli, we found no evidence of discrimination between test grammars based on their consistency with the familiarization grammar, and no age-related differences between six- and nine-month-olds. In Experiment 1, we also tested a group of infants using birdsong, with no difference in results. Instead, what we found in both experiments was an overall preference for adjacent repetition items (XXY) during the test, independent of training grammar. In Experiment 3, in an attempted close replication of the original Marcus et al. experiment, we found that seven-months-old performed in a similar manner, preferring the XYY grammar during the test, regardless of whether they were familiarized with XYY or XYX. In both cases, a Bayesian analysis strengthened our findings and allowed us to conclude that infants across the three experiments showed strong evidence of having an adjacent repetition preference.

In chapter 3, we turned to the visual domain. In that chapter we attempted to extend on a study from Saffran et al. (2007), who had shown that infants at seven months were able to generalize Marcus rules in the visual domain when the grammars were composed of pictures of dogs or of cats of different breeds. While their conclusion was that success in this experiment was due to infants'

familiarity with the stimuli and their ability to categorize it, we hypothesized that infants may have been successful in that task due to the fact that they probably had a lexical representation for dogs or cats (or similar animate objects) at this age. This would entail that rule learning is still facilitated by linguistic cues. We tested whether there would be a difference between learning Marcus grammars in the visual domain when the visual stimuli consisted of (1) items that 12- to14-month-old infants knew the names for, or (2) of items that were familiar to them but not lexically specified, or (3) of nonsense objects. While we found that they looked longer at stimuli consisting of items they knew the names for, this did not translate to success in learning about the underlying grammatical rules in all but boys familiarized with XYY in the Known condition. This limited finding should be interpreted with caution. While we have theoretical and empirical reasons to trust that there would be differences between groups related to their sex (e.g., Shi, 2007; Mueller et al., 2012; Lany & Gómez, 2008; ter Haar & Levelt, 2018) and training grammar (e.g., S. P. Johnson et al., 2009; Endress et al., 2009), and although the effect was strong, the number of participants was low and the effect should be replicated to confirm the finding. Nevertheless, the work in this chapter suggests that item familiarity in itself does not guarantee generalization of the Marcus rules to the visual domain. A factor that could have contributed to the success in Saffran et al. (2007), which could be tested in future work, is the fact that dogs and cats are animate, which may make them more interesting stimuli for infants of this age to learn from. Neither the abstract shapes in Johnson et al. (2007), nor the pictured items in our experiments were animate.

In chapter 4, we looked at production in infants, seeking to better understand the types of patterns present in their early babbling phase as a means to contextualize the ability (or inability) to learn certain patterns in perceptual tasks. Analysis of more than 60,000 utterances in nine languages from children in their first two years of life showed that contrary to established theories of language production development, completely variegated utterances (i.e., following patterns like XY, XYZ, etc., where the variables represent CV syllables) were present from the earliest utterances and were the most frequent type of utterances across the languages. Of the utterances containing repetition, those containing immediate repetition (or reduplication; XXY and XYY) were far more frequent than those containing non-adjacent repetition (XYX). These results may seem at first glance to contradict our experimental findings, with variegated productions being the most produced, but immediately repeating ones attracting the most attention, i.e., longer looking times. However, repetition has been found to be a learning tool in production, and as such, it follows that repeated utterances attract attention. Utterances such as XYX may be in an unhappy medium, preferred neither in perception nor in production due to the memory constraints necessary both to perceive and to produce them.

Chapters 5 and 6 focused on adult rule learning, using Marcus grammars and the more complex Lindenmayer grammars respectively. Studying adults

allowed us to gain insights into factors potentially influencing rule learning that we could not get from infant participants. Our adult studies allowed us to manipulate the task and to look at both implicit and explicit factors. Chapter 5 revealed that Marcus-type rule learning does not proceed unimpeded in adults; they, like infants, can only learn under specific circumstances. Adults could not generalize the simple rules when instructions were too vague. Instead, they benefitted from explicit instructions, from an active learning task with feedback, or from specific test stimuli that could only be processed when generalizing, directing them to the level at which they should attend. On the other hand, chapter 6 revealed that adult participants are able to pick up on regularities in far more complex stimuli, involving hierarchical structures. In both an identification task (Yes/No) and an identification-discrimination task (2AFC), participants showed sensitivity to an L-system grammar instantiated with different types of drum sounds.

The sections that follow will more explicitly expand on these findings by categorizing them according to the themes identified in chapter 1.

7.2 Domain specificity

While the results of chapters 2 and 3 do show a difficulty to learn the Marcus-type rules, they also lend support to the notion that at least some of the biases and processing mechanisms involved in learning rules are not specific to language. In chapter 2, we tested infants in the auditory domain, using both speech and birdsong stimuli and found no differences between the performance in the two domains. While no general rule learning was found in either condition, infants in both conditions showed the same type of preference in the test phase, namely for test stimuli containing adjacent repetitions. This finding indicates that independent of domain, a bias for repetitions underlies infants' general processing of auditory structures. In the visual rule learning tasks of chapter 3, we found that at least a subset of participants were able to generalize the learned rule, namely male infants familiarized with XYY grammars in the condition in which they had lexical representations for the items in the pictures. While this result is tentative and should be replicated, it adds support to the idea that infants are able to learn rules from more domain-general visual stimuli, but perhaps only when facilitated by a repetition pattern and the availability of a lexical representation of the pictured objects.

In chapter 6, where we tested adult participants using drum sounds rather than speech, participants showed sensitivity to the L-system grammar as compared to a foil grammar with similar surface properties. Although we did not explicitly test domain specificity within our task, comparing our results to those found by Shirley in the speech domain can give us an indication of adults' ability to learn L-systems within and outside the speech domain. Shirley's results were more robust; we can compare the mean correctness scores of participants in her study using speech stimuli and ours using drum sounds in the same

type of task (2AFC), and see that participants in her speech task performed with an accuracy rate of nearly 20% higher than participants in our chapter 6. The discrepancy in results adds evidence to the claim that speech has a facilitatory role in aiding rule learning. Nevertheless, our participants did perform above chance level, also suggesting that they did become sensitive to the grammars. One explanation for the differences between chapters 1 through 5 and of chapter 6 may be that with simpler grammars, few differences can be discerned between domains; with more complex grammars, speech may hold an advantage - unsurprising since this is the domain in which humans do perform complex rule learning and production.

In addition, it is relevant to return to the original AGL studies (such as those performed by Reber, 1967, 1976) in discussing this notion of domain specificity or generality. Note that while these first studies on AGL did not focus on domain generality *per se* and were meant to illuminate mechanisms of language acquisition, they were, in fact, instantiated not using spoken or signed language, but with visually-presented letters on a screen whose order of presentation was generated by an underlying grammar. This method of presentation already removes domain-specific learning mechanisms from consideration, as the linguistic value of a string of orthographically presented consonants is questionable. The use of this type of stimuli continues in many AGL tasks with adults (e.g., Kinder & Assmann, 2000; Johnstone & Shanks, 1999; see also Perruchet & Pacton, 2006 for an overview).

The potential to learn patterns across domains corroborates cross-species findings as well. Birds, for example, are able to pick up regularities at some level in auditory stimuli: zebra finches at the level of the sequential ordering of items within a triad, budgerigars even at the more abstract level of rules (Spierings & ten Cate, 2016). Rule learning abilities that humans use for language acquisition may thus have roots in these more domain-general and species-general abilities to process and categorize information. On the other hand, similar outcomes do not necessarily indicate similar processes (Milne, Wilson, & Christiansen, 2018), and how rule learning proceeds online in each domain should be further explained. Nevertheless, there may be shared biases that aid rule learning across domains. One of these mechanisms from which learning seems to benefit is the perceptual sensitivity to repetitions, which has been found across ages, domains, and species (Endress & Bonatti, 2007; Endress et al., 2009; van Heijningen et al., 2013; Chen et al., 2015).

7.3 Repetitions

In chapter 2, and to some extent also in chapter 3, we found evidence for an asymmetry in performance after exposure to the types of Marcus grammars containing immediate repetitions (XYY and XXY) and those containing non-adjacent repetitions (XYX). Chapter 4, which looked at the patterns found in infants' earliest utterances, attempts to link these perception results with

what infants do around the same period in production. The main finding was that infants produce significantly more variegated utterances (like XYZ) than non-variegated ones (like XYY, XXY or XYX), and consistently so throughout their first year. The secondary finding is that they produce more non-variegated utterances containing immediate repetitions (XYY, XXY) than non-adjacent repetitions (XYX). How can we reconcile these findings and how do they help inform us about what is happening in our perception experiments?

Repetition in perception has long been identified as an important factor supporting learning. Gómez et al. (2000) identify repetition as the basis of the ability to generalize (or transfer) rules. While Gomez et al. addressed learning and generalization of a finite state grammar, the finding is relevant for rule learning in general. Repetition is highly salient for humans from the beginning of life (Gervain et al., 2008), and is relevant for learning across domains (S. P. Johnson et al., 2009).

The role of repetition in production is linked to learning too. Based also on corpus results, Lipkind et al. (2013) argued that repetitions in production serve a memory function, helping to solidify the memory of a newly-acquired syllable in the infant's inventory. Experimental work from Fagan (2015) also supports this notion of the role of repetition in infants' productions. Observations of normally-hearing infants and those learning language after receiving a cochlear implant indicated that repetitions in production emerged as vocal exercises, strengthening the perceptual memory of new phonemes and syllables.

Returning to rule learning, we can see that if syllable repetition is a learning tool for infants, both in perception and in production (and its interaction) it follows that the rules containing immediate repetition in the Marcus et al. (1999) paradigm would facilitate learning in that task too.

Where does this leave seemingly non-preferred utterances such as XYX containing non-adjacent repetitions? Gervain et al. (2008) found asymmetries between discrimination abilities of newborns tasked with discriminating XXY vs. XYZ and XYX vs. XYZ. While the former produce evidence of discrimination abilities, the latter did not. The fact that in perception XYX and XYZ are not discriminated, and that in production XYX is so rarely produced in comparison to all other utterances, may indicate these are perceived (and produced, perhaps accidentally) as a subset of the variegated XYZ.

Turning to adults, chapter 5 and 6, show that repetition is not the only basis for abstraction and that perhaps adults are less sensitive to repetition than infants. In chapter 5, we found no evidence of asymmetry in performance between adults familiarized with repetition- versus non-repetition grammars. In addition, in chapter 6, using a far more complex grammar, we found that participants were able to correctly identify and discriminate the grammatical from ungrammatical strings even though both streams shared a similar surface distribution of immediate repetition bigrams. In order to discriminate this similar non-grammatical foil from that following an L-system, adults in this task must have learned far more about the grammar than the fact that it includes

repetition. Even if participants in this task were to use a strategy that has an expectation of repetition, such a strategy would still require an anticipation of where the repetition can and cannot be in order to correctly identify and discriminate the correct grammar from the incorrect one. Therefore, an implicit understanding of the structure must have been formed. Thus, while repetition may be helpful and salient throughout development, and may aid in learning, the bias for simple repetition grammars found in our infants can be overcome. In the case of our adults, this can presumably come as a result of a number of cues, several of which are discussed further in section 7.5.

7.4 Familiarity

The work in chapter 3 attempted to clarify the level at which stimuli had to be familiar in order to facilitate rule learning. Saffran et al. (2007) claimed that their success in showing rule learning in the visual domain and the failures and asymmetries found in other visual rule learning tasks (e.g., S. P. Johnson et al., 2009) was a result of the level of the infants' familiarity with the objects presented (high for dogs in Saffran et al., 2007 and low for shapes in S. P. Johnson et al., 2009). However, as mentioned in chapter 1, familiarity may be at the level of the pictured item itself or at the level of the item's lexical representation. In chapter 3, while we did find that infants were significantly more attentive to items with which they had lexical familiarity, this behavior did not translate to better overall learning. While we found some evidence of learning in the condition in which participants knew the picture labels, overall performance in that condition was not strongly indicative of learning, since it occurred in only one sex (male), trained with one grammar (XYY). This result needs to be replicated with more participants in order to draw stronger conclusions. What can be said, however, is that familiarity at the item level did not seem to facilitate generalization; compared to nonsense items, which could not have been familiar, infants did not perform better with familiar stimuli.

The findings of chapter 2 also support the conclusion that item familiarity does not make a clear difference in learning performance. Infants in this case performed similarly, whether exposed to speech or to birdsong stimuli, the latter of which they would not be highly familiar with.

The argument for familiarity being beneficial for generalization, therefore does not find support in the work presented here.

7.5 Ecological validity

While it is surprising that infants struggle to learn simple Marcus-type grammars when they can learn the complex grammar of natural language, it is of note that the tasks infants are asked to perform in rule learning studies are not representative of the natural language or natural learning environments in

which either infants or adults actually learn. While infants are able to learn a great deal implicitly, their learning environment is not only implicit but, also includes explicit feedback, repetitions of words and utterances, exaggerated intonation contours that direct attention to specific structures, and positive (verbal and non-verbal) reinforcement from interlocutors. Adult learning of natural language (e.g., a foreign language), on the other hand, is often not implicit but, depending on the learner, often benefits from varied formal instruction and feedback. While AGL tasks do not pretend to be representative of natural learning circumstances, the end goal of performing such experiments is to be able to tap into learning abilities used in natural grammar learning. Yet if we show multiple failures of "simple" learning, we fail at that end goal, and the ecological validity of such tasks can be questioned on multiple fronts.

7.5.1 Variability

One way in which we attempted to make our rule learning tasks more ecologically valid was to manipulate the amount of variability in the input. Previous Marcus-type rule learning experiments have typically been limited with respect to the phonological variability of the auditory stimuli. In Experiments 1 and 2 of chapter 2 with infants, and in chapter 5 with adults, we used a set of syllables that were phonologically balanced to be more representative of the natural language to which participants are normally exposed. When comparing the results of Experiments 1 and 2 of chapter 2 with those of Experiment 3, where infants heard the low-variability Marcus et al. stimuli, we found no difference in performance. Increasing the variability of the phonemes and syllables to make them more naturalistic did not produce any advantage in learning. In addition, in chapter 5, we manipulated the amount of variability in the stream by varying the number of triads presented to the participant: either 15 triplets repeated three times or three triplets repeated 15 times. We found that while learning was still successful when participants were exposed to three triplets, a boost was seen as a result of being exposed to 15 phonologically different triplets.

However, the fact remains that learning in adults was indeed still possible beyond chance level with a low-variability stream, and infants did not receive any boost in performance whether they were exposed to our phonologically well-balanced set of syllables and triplets versus those used in Marcus et al.'s (1999) original study. Together these results from infants and adults indicate that variability may play a role in learnability, but it is probably not a key factor. This finding fits well with results from Gerken (2010; 2015) that indicate that infants are able to learn a rule from a handful of stimuli, or even from just one stimulus, given the correct circumstances.

7.5.2 Task and instruction

In chapter 5, we systematically varied task and instruction in a series of Marcus-type rule learning tasks with adults. When presented with vague instructions,

resulting in a passive learning task in which they were unaware of what level of the stimulus stream they should attend to, adults struggled to generalize the underlying rule to novel stimuli. This passive learning task was arguably the most similar to the experimental conditions that infants undergo, in which no instruction can be given and nothing in particular directs their attention to anything specific about the stimuli. On the other hand, adults succeeded in the tasks when their attention was directed to the relevant level of information (the grammar) by manipulations in the instruction or in the type of task presented. An active, Go-Left/Go-Right task, analogous to the Go/No-Go task animals typically perform, yielded highly successful results.

However, in our L-systems experiments, task did not seem to make a difference in participants' ability to discriminate grammatical from non-grammatical strings. In this case, there were no differences between a 2AFC task and a Yes/No tasks overall : participants in both cases identified the correct strings at a rate higher than would be expected by chance. In this case, instructions were more explicit than those provided during some of the experiments in chapter 5, calling attention to the fact that a rhythmic pattern would be heard and need to be classified later in test. Yet the tasks were similar to those in chapter 5, consisting of an identification task and a discrimination task. Participants here performed above chance in both tasks. And since each participant performed both tasks, we were able to study the influence of the first on the second task. Participants performed better on the 2AFC task when it was performed first, although there was no effect of order of tasks on performance in the Yes/No task. These task differences give us an indication that identification tasks may be more robust while performance on discrimination tasks may become weaker over time after exposure to a different type of task.

Like the adults in the passive learning task, infants struggled to learn simple grammars as well; whether their difficulty was a result of the task or not is still an empirical question. However when comparing the complex structures that infants can make sense of in the natural world with what they are unable to do in the lab, the question arises of whether we give infants a fair chance to show us what they can do. In an experimental environment, as we have seen above, adults need instruction or feedback to learn. Animals also receive feedback in AGL tasks, both positive (e.g., a food reward) and negative (e.g., a short period of darkness), and they are typically trained until they reach a certain criterion of performance (e.g., Spierings & ten Cate, 2016). Infants in the real world also receive feedback from their caregivers, both verbal and non-verbal. While short passive exposure tasks might be sufficient for simple discrimination studies (e.g., Maye et al., 2002; Maye, Weiss, & Aslin, 2008), the lack of feedback in the rule learning experiments might thus impede infants' learning, not allowing them to show their full potential in grammar learning tasks. Alternatively, if infants are able to learn the grammar of a natural language implicitly, but not a Marcus grammar, this could imply that the learning of Marcus grammars actually does not model natural language learning very well.

In addition, there is ambiguity in interpreting infant looking times in the sense that we are not always sure what it means with respect to the individual infant, the stimuli, and the task as a whole (Kidd et al., 2010). There are many hidden variables that can contribute to this one dependent measure (Aslin, 2007). The preference that we measure is a proxy for discrimination, but an infant may be able to perceive a difference between two stimuli, while preferring neither of the two. This may be why no preference between XYX and XXY grammars was found when pilot-testing a pure preference task, presenting infants with either XYX or XXY stimuli without a preceding familiarization phase (see discussion of chapter 2). Their bias for the repetition grammar thus emerges with respect to what they heard during familiarization. Research has shown that infants' preferences are volatile, and that the stimuli they are provided with must contain the exact right amount of information to draw and hold their attention (the "Goldilocks effect"; Kidd et al., 2010).

7.5.3 Grammar complexity

In chapters 1 and 2, we discuss the notion that while immediate repetition is found throughout the world's languages at the word and phrase level, the XYX rule is not typically found in natural language. This lack of naturalness in the stimuli is reflected in the fact that, as mentioned above, this grammar was not learned or preferred by infants in chapters 2 and 3, nor was it readily learned by adults in chapter 5 without certain task, instruction, or attention conditions being met. This leads us to the conclusion that these grammars are not well suited for our greater understanding of rule learning in infants or in adults. And it should not be surprising, since they do not reflect much about the task that we are attempting to explore.

In chapter 6 we experimented with a far more complex grammar, generated by simple recursive rules. The work in this chapter gives us an indication that we may be more successful in better understanding how participants learn about language and structure by employing artificial grammars that are more ecologically valid. The L-systems that were explored in chapter 6 have not been widely used in AGL work up until now. The task is difficult and outside of the linguistic domain, making use of drum sounds rather than speech. Yet participants do show some evidence of sensitivity to the difference between grammatical and ungrammatical strings during test at a rate higher than would be expected by chance, in both an identification and a discrimination-identification task. While these results may be explained by a more superficial error detection mechanism rather than a process of learning about the underlying structure of the grammar as a whole, and while they are not evidence of generalization to novel stimuli or to a novel domain, they do indicate that there is some mechanism within participants that presumably allows them to pick up on some level of regularity even in this complex stimuli. While more exploration is clearly needed, it is likely the route of more complexity that will provide us with more interesting and applicable results. The fact that this grammar has rewrite rules

that allow for recursive patterns brings it much closer to natural languages than what is used in the simple rule learning experiments typically used in infant research (e.g., Marcus et al., 1999; Gómez, 2002). Future experiments with these grammars may attempt to use linguistic stimuli to study whether infants are as sensitive to these complex grammars as adults were in the work of Shirley (2014).

7.6 Conclusion

The work presented in this thesis has shown that generalization is not easy, and simple rule learning is not so simple. Work on Marcus-type rule learning has proved difficult to replicate, and failures to replicate are, in turn, difficult to interpret. Perhaps these grammars are not the best model grammars to use for understanding rule learning in natural language, or for understanding how or why humans pick up on regularities in the input. If these experiments are difficult to replicate, if they have small effect sizes when they do succeed, or if successes only occur under specific circumstances, how can we identify the cognitive mechanisms humans use to learn the structure of natural language? AGL should be a reliable tool for investigating how we learn natural language, and when infants fail to learn in an AGL experiment, we want to be sure it is because of their cognitive mechanism and not because of the method we are using.

Future studies into rule learning capabilities should address several of the issues raised in this thesis. First, artificial grammars used in these studies can and should be more complex and naturalistic. In addition, to get away from some of the methodological pitfalls identified above, future work could focus more on online measurements rather than the indirect measures such as the looking times typically collected from infants and the Yes/No-type responses typically collected from adults. In addition to EEG and NIRS studies, this could include online tasks involving serial reaction times (SRT), pupil dilation, or recall tasks (e.g., statistically-induced chunking recall, Isbilen, McCauley, Kidd, & Christiansen, 2017). Online measurements allow us to gain further insights into what participants find to be informative or surprising stimuli, and into the time-course of processing, which would inform us about the difficulty of the stimuli or task.

Finally, future work should include Bayesian analyses for better understanding of the data, particularly for understanding null results. Along with tools such as meta-analyses of experimental data and corpus analyses of production data, such analyses of null results and failures can help reshape the picture we have of how learning occurs in infants as well as in adults.

References

Alhama, R. G. (2017). *Computational modeling of artificial language learning: Retention, recognition and recurrence.* University of Amsterdam.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30*(5), 666–677.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Beckers, G. J., Bolhuis, J. J., Okanoya, K., & Berwick, R. C. (2012, feb). Birdsong neurolinguistics. *NeuroReport*, *23*(3), 139–145.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–8.

Bernard, C., & Gervain, J. (2012). Prosodic cues to word order: What level of representation? *Frontiers in Psychology*, *3*(OCT), 451.

Boersma, P., & Weenink, D. (2007). Praat: doing phonetics by computer (Version 4.5.)[Computer program]. *Retrieved from http://www.praat.org/*, *5*(9/10), 341–345.

Braine, M. D. S. (1987). *What is Learned in Acquiring Word Classes - A Step Toward an Acquisition Theory.*

Bunge, S. A., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, *15*(3), 118–121.

Chen, J., & ten Cate, C. (2015). Zebra finches can use positional and transitional cues to distinguish vocal element strings. *Behavioural Processes*, *117*, 29–34.

Chen, J., van Rossum, D., & ten Cate, C. (2015). Artificial grammar learning in zebra finches and human adults: XYX versus XXY. *Animal Cognition*, *18*(1), 151–164.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2000). A Connectionist Single-Mechanism Account of Rule-Like Behavior in Infancy. *Proceedings*

of the 22nd Meeting of the Cognitive Science Society, *1*(Experiment 1), 83–88.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science*, *1*(2), 215–219.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*(12), 997–1003.

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities : Pitting abstract against stimulus-specific representations. *Psychological Science*, *17*(10), 905–912.

Cooper, R. P., & Aslin, R. N. (1990). Preference for Infant-Directed Speech in the First Month After Birth. *Child Development*, *61*(5), 1584–1595.

Corballis, M. C. (2007). Recursion, language, and starlings. *Cognitive science*, *31*(4), 697–704.

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1119–1130.

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2015). Statistical treatment of looking-time data. *Developmental Psychology*(2015), 1–16.

Dale, R., & Christiansen, M. H. (2004). Active and passive statistical learning: Exploring the role of feedback in artificial grammar learning and language. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 26).

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1199–1211.

de Boysson-Bardies, B. (2001). *How language comes to children: From birth to two years*. MIT Press.

de Boysson-Bardies, B., & Vihman, M. M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language*, *67*(2), 297–319.

Demuth, K., Culbertson, J., & Alter, J. (2006). and Coda Licensing in the Early Acquisition of English.

Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of french determiners. *Journal of Child Language*, *35*(1), 99–127.

DePaolis, R. A., Vihman, M. M., & Nakai, S. (2013). The influence of babbling patterns on the processing of speech. *Infant Behavior and Development*, *36*(4), 642–649.

de Vries, M., Christiansen, M., & Petersson, K. M. (2011). Learning recursion: Multiple nested and crossed dependencies. *Biolinguistics*, *5*(1-2), 010–035.

de Vries, M., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded

hierarchical structures. *Cognition*, *107*(2), 763–774.

Edmunds, C., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, *68*(6), 1203–1222.

ELAN Version 5.0.0. (2017). Nijmegen: Max Planck Institute for Psycholinguistics.

Endress, A. D. (2013). Bayesian learning and the psychology of rule induction. *Cognition*, *127*(2), 159–176.

Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*(2), 247–299.

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition*, *105*(3), 577–614.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*(3), 406–419.

E-Prime 2.0, Psychology Software Tools, Inc. (2012).

Fagan, M. K. (2015). Why repetition? Repetitive babbling, auditory feedback, and cochlear implantation. *Journal of Experimental Child Psychology*, *137*, 125–136.

Ferguson, B., & Lew-Williams, C. (2016). Communicative signals support abstract rule learning by 7-month-old infants. *Nature Publishing Group*(May), 1–7.

Ferguson, B., & Waxman, S. R. (2015). Visual abstract rule learning by 3- and 4-month-old infants. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 692–697.

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*(3), 329–364.

Fitch, W. T., & Friederici, A. D. (2012, jul). Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *367*(1598), 1933–55.

Fitch, W. T., & Hauser, M. D. M. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*(5656), 377.

Forkstam, C., Elwér, Å., Ingvar, M., & Petersson, K. M. (2008). Instruction effects in implicit artificial grammar learning: A preference for grammaticality. *Brain Research*, *1221*, 80–92.

Frank, M. C. (2013). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). *Cognition*, *128*(3), 417–423.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science*, *12*(4), 504–509.

Frank, M. C., & Tenenbaum, J. B. (2011, sep). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

Freitas, M. J. (1997). *Aquisição da estrutura silábica do português europeu.*

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: functional localization and structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(7), 2458–63.

Geambașu, A., Ravignani, A., & Levelt, C. C. (2016). Preliminary experiments on human sensitivity to rhythmic structure in a grammar with recursive self-similarity. *Frontiers in Neuroscience*, *10*, 281.

Geambașu, A., Scheel, M., & Levelt, C. C. (2016). Cross-linguistic patterns in infant babbling. In *Proceedings of the 40th annual Boston University Conference on Language Development.* Cascadilla Press.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical Learning of Adjacent and Non-Adjacent Dependencies among Non-Linguistic Sounds. *Psychonomic Bulletin & Review*, *16*(3), 486–490.

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2010). NIH Public Access. , *16*(3), 486–490.

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–74.

Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, *115*(2), 362–366.

Gerken, L., & Bollt, A. (2008). Three Exemplars Allow at Least Some Linguistic Generalizations: Implications for Generalization Mechanisms and Constraints. *Language Learning and Development*, *4*(3), 228–248.

Gerken, L., Dawson, C., Chatila, R., & Tenenbaum, J. (2015). Surprise! Infants consider possible bases of generalization for a single input example. *Developmental Science*, *18*(1), 80–89.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(2), 249–268.

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 14222–14227.

Gervain, J., & Werker, J. F. (2013). Learning non-adjacent regularities at age 0 ; 7. *Journal of Child Language*, *40*(4), 860–72.

Gómez, R. L. (1997). Transfer and Complexity in Articial Grammar Learning. *Cognitive Psychology*, *33*, 154–207.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological science : a journal of the American Psychological Society / APS*, *13*(5), 431–436.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds

leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–35.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, *4*(5), 178–186.

Gómez, R. L., Gerken, L., & Schvaneveldt, R. W. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition*, *28*(2), 253–263.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Hay, J. S. F., & Diehl, R. L. (2007). Perception of rhythmic grouping: testing the iambic/trochaic law. *Perception & psychophysics*, *69*(1), 113–122.

Hochmann, J. R., Azadpour, M., & Mehler, J. (2008). Do Humans Really Learn A^nB^n Artificial Grammars From Exemplars? *Cognitive Science*, *32*, 1021–1036.

Hoff, E. (2008). *Language development*. Cengage Learning.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: a novel chunk-based measure of statistical learning. In *the 39th Annual Conference of the Cognitive Science Society (CogSci 2017)* (pp. 564–569).

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291.

JASP Team. (2017). *JASP*.

Jeffreys, H. (1961). The theory of probability. *International series of monographs on physics*.

Johansson, T. (2009). Strengthening the case for stimulus-specificity in artificial grammar learning. *Exp Psychol*, *56*(3), 188–197.

Johnson, E. K., & Zamuner, T. (2010). Using infant and toddler testing methods in language acquisition research. *Experimental Methods in Language Acquisition Research*, 73–94.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract Rule Learning for Visual Sequences in 8- and 11-Month-Olds. *Infancy*, *14*(1), 2–18.

Johnstone, T., & Shanks, D. R. (1999). Two Mechanisms in Implicit Artificial Grammar Learning ? *Cognition*, *25*(2), 524–531.

Johnstone, T., & Shanks, D. R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology*, *42*(1), 61–112.

Junge, C., Cutler, A., & Hagoort, P. (2012). Electrophysiological evidence of early word learning. *Neuropsychologia*, *50*(14), 3702–3712.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.

Jusczyk, P. W., & Luce, P. (1994). Infants' Sensitivity to Phonotactic Patterns

in the Native Language. *Journal of Memory and Language*, *33*(5), 630–645.

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*(1), 111–116.

Kern, S., Davis, B., & Zink, I. (2009). From babbling to first words in four languages: Common trends across languages and individual differences.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2010). The Goldilocks Effect: Infants' preference for stimuli that are neither too predictable nor too surprising. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 2476–2481.

Kinder, A., & Assmann, A. (2000). Learning artificial grammars: no evidence for the acquisition of rules. *Memory & Cognition*, *28*(8), 1321–1332.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of experimental psychology. Learning, memory, and cognition*, *22*(1), 169–181.

Koopmans-van Beinum, F. (1980). *Vowel contrast reduction: An acoustic and perceptual study of Dutch vowels in various speech conditions.*

Koopmans-Van Beinum, F. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In *Precursors of early speech* (pp. 37–50). Springer.

Kovács, Á. M., & Endress, A. D. (2014). Hierarchical processing in seven-month-old infants. *Infancy*, *19*(4), 409–425.

Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6556–60.

Kuhn, G., & Dienes, Z. (2005). Implicit Learning of Nonlocal Musical Rules: Implicitly Learning More Than Chunks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1417–1432.

Kuhn, G., & Dienes, Z. (2006). Differences in the types of musical regularity learnt in incidental- and intentional-learning conditions. *The Quarterly Journal of Experimental Psychology*, *59*(10), 1725–1744.

Lai, J., & Poletiek, F. H. (2013). How "small" is "starting small" for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology*, *25*(4), 423–435.

Lany, J., & Gómez, R. L. (2008). Twelve-Month-Old Infants Benefit From Prior Experience in Statistical Learning. *Psychological Science*, *19*(12), 1247–1252.

Levelt, C. C. (1994). *On the acquisition of a place feature representation.* Dordrecht.

Levelt, W. (2013). *A history of psycholinguistics: The pre-Chomskyan era.* Oxford University Press.

Lindenmayer, A. (1968). Mathematical Models for Cellular Interactions in

Development I. Filaments with One-sided Inputs. *Journal of Theoretical Biology*, *18*, 280–299.

Lindenmayer, A., & Rozenberg, G. (1972). Developmental systems and languages. In *STOC '72 proceedings of the fourth annual ACM symposium on theory of computing* (pp. 214–221).

Lipkind, D., Marcus, G. F., Bemis, D. K., Sasahara, K., Jacoby, N., Takahasi, M., ... Tchernichovski, O. (2013). Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature*, *498*(7452), 104–8.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662.

Majorano, M., Vihman, M. M., & DePaolis, R. A. (2014). The relationship between infants? production experience and their processing of speech. *Language Learning and Development*, *10*(2), 179–204.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. *Psychological Science*, *18*(5), 387–391.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Martins, M. D. (2012). Distinctive signatures of recursion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 2055–2064.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101—-11.

Mealor, A. D., & Dienes, Z. (2013). Explicit feedback maintains implicit knowledge. *Consciousness and Cognition*, *22*(3), 822–832.

Mehler, J., Nespor, M., & Peña, M. (2008). What Infants Know and What They have to Learn about Language. *European Review*, *16*(04), 429.

Milne, A. E., Wilson, B., & Christiansen, M. H. (2018). Structured sequence learning across sensory modalities in humans and nonhuman primates. *Current Opinion in Behavioral Sciences*, *21*, 39–48.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Mueller, J. L., Friederici, A. D., & Mannel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences*, *109*(39), 15953–15958.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Newport, E. L., Hauser, M. D., Spaepen, G., & Aslin, R. N. (2004). Learning at a distance II. Statistical learning of non-adjacent dependencies in a non-human primate. *Cognitive Psychology*, *49*(2), 85–117.

Oller, D., Yeni-Komshian, G., Kavanaugh, J., & Ferguson, C. (1980). Child phonology.

Onnis, L., Christiansen, M. H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 25).

Opitz, B., Ferdinand, N. K., & Mecklinger, A. (2011). Timing matters: the impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience*, *5*(February), 8.

Ota, M. (2003). *The development of prosodic structure in early words: Continuity, divergence, and change* (Vol. 34). John Benjamins Publishing.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science (New York, N.Y.)*, *298*(5593), 604–607.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*(3), 264–275.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–238.

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning nonadjacent dependencies: no need for algebraic-like computations. *Journal of Experimental Psychology: General*, *133*(4), 573–583.

Petersson, K.-M., Folia, V., & Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language*, *120*(2), 83–95.

Poletiek, F. H., Fitz, H., & Bocanegra, B. R. (2016). What baboons can (not) tell us about natural language grammars. *Cognition*, *151*(2016), 108–112.

Prusinkiewicz, P., & Lindenmayer, A. (1990). *The Algorithmic Beauty of Plants (The Virtual Laboratory)*.

Rabagliati, H., Senghas, A., Johnson, S. P., & Marcus, G. F. (2012). Infant rule learning: Advantage language, or advantage speech? *PLoS ONE*, *7*(7), e40517.

Ravignani, A., Westphal-Fitch, G., Aust, U., Schlumpp, M. M., & Fitch, W. T. (2015). More than one way to see it: Individual heuristics in avian visual computation. *Cognition*, *143*, 13–24.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning*

and Memory, *2*(1), 88–94.

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology*, *118*(3), 219–235.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology Human Learning and Memory*, *6*(5), 492–502.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science*, *37*(7), 1290–1320.

Rose, Y., & MacWhinney, B. (2014). The PhonBank project. In *The Oxford handbook of corpus phonology.*

Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing phon: A software solution for the study of phonological acquisition. In *Proceedings of Boston University Conference on Language Development* (Vol. 2006, p. 489).

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Roug, L., Landberg, I., & Lundberg, L. (1989). Phonetic development in early infancy : a study of four Swedish children during the first eighteen months of life. *Journal of Child Language*, *16*, 19–40.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680.

Saffran, J. R., & Wilson, D. P. (2003). From Syllables to Syntax : Multilevel Statistical Learning by 12-Month-Old Infants. *Infancy*, *4*(2), 273–284.

Samal, A., Peterson, B., & Holliday, D. J. (1994). Recognizing plants using stochastic L-systems. In *Proceedings - International Conference on Image Processing, ICIP* (Vol. 1, pp. 183–187).

Schlichting, L., & Lutje Spelberg, H. (2009). Lexilijsten en Lexiconlijsten, oudervragenlijsten voor de taalontwikkeling. *Logopedie en Foniatrie*, *6*, 190–195.

Shi, R. (2007). Infants' Recognition of Function Words in Continuous Speech. *ICPhS XVI*(August), 1541–1544.

Shirley, E. J. (2014). *Representing and remembering Lindenmayer grammars.* University of Reading.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language*

Resources and Evaluation (LREC'08), 816–820.

Smith, B. L., Brown-Sweeney, S., & Stoel-Gammon, C. (1989). A quantitative analysis of reduplicated and variegated babbling. *First Language*, *9*(6), 175–189.

Spierings, M. J. (2016). *The music of language in zebra finches and budgerigars.* Leiden University.

Spierings, M. J., & ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. *Proceedings of the National Academy of Sciences*, *113*(27), E3977–E3984.

Stark, R. E. (1980). Stages of speech development in the first year of life. In *Child phonology* (pp. 73–92). Elsevier.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Sun, F., Hoshi-Shiba, R., Abla, D., & Okanoya, K. (2012). Neural correlates of abstract rule learning: An event-related potential study. *Neuropsychologia*, *50*(11), 2617–2624.

ter Haar, S. M., & Levelt, C. C. (2018). Disentangling attention for frequency and phonological markedness in 9-and 12-month-old infants. *Language Learning and Development*, 1–18.

Thiessen, E. D. (2012). Effects of Inter- and Intra-modal Redundancy on Infants' Rule Learning. *Language Learning and Development*, *8*(3), 197–214.

Uddén, J., Ingvar, M., Hagoort, P., & Petersson, K. M. (2012). Implicit Acquisition of Grammars With Crossed and Nested Non-Adjacent Dependencies: Investigating the Push-Down Stack Model. *Cognitive Science*, *36*(2012), 1078–1101.

Uriagereka, J., Reggia, J. A., & Wilkinson, G. S. (2013, jan). A framework for the comparative study of language. *Evolutionary psychology : an international journal of evolutionary approaches to psychology and behavior*, *11*(3), 470–92.

van der Kant, A. (2015). *Neural correlates of vocal learning in songbirds and humans.*

van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B., & ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal Cognition*, *16*(2), 165–75.

van Heijningen, C. A. A., de Visser, J., Zuidema, W., & ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20538–20543.

Vihman, M. M. (1996). *Phonological development: The origins of language in the child.* Blackwell Publishing.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*,

25(3), 169–176.

Whittlesea, B. W., & Dorken, M. D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. *Journal of Experimental Psychology: General*, *122*(2), 227–248.

Zimmerer, V. C., Cowell, P. E., & Varley, R. a. (2011). Individual behavior in learning of an artificial grammar. *Memory & Cognition*, *39*(3), 491–501.

Summary

The work presented in this thesis is part of an interdisciplinary project aimed at understanding the cognitive roots of language acquisition. The goal of this work was to understand the underpinnings of grammatical rule learning by humans: whether our ability to learn our complex human languages is purely specific to us, whether the building blocks of learning language is a general ability that can be applied in other domains (e.g., the auditory domain in general, the visual domain), and how the way a learning task is presented to participants may affect their ability to learn from the input they receive. In the larger project, our collaborators aimed to understand how human abilities might compare to those of other species, namely zebra finches and budgerigars, and to create a comprehensive computational model that can help explain the principles underlying learning of segments of information and the rules that dictate their structure.

This thesis is a collection of papers tackling rule learning from various perspectives, in both infants and adults. Using the technique of creating highly controlled artificial "languages" (referred to in general as an Artificial Grammar Learning paradigm), we were able to gain some insights into learning processes of young infants within their first year of life and of adults, mostly of university age.

Following a general introduction, chapter 2 reports on an attempt to replicate an often-cited set of experiments from Marcus and colleagues (1999) showing that infants as young as seven months old can quickly learn simple patterns following either an XYX, XXY, or XYY rule and recognize those rules when they are presented with speech sounds that were not heard in their learning phase of the experiment. This study has been cited time and again as providing evidence that infants are able to generalize, or transfer, a learned rule to novel information. Yet our large-scale attempt at extending on, and eventually replicating closely, the Marcus et al. (1999) study, was unsuccessful in replicating the original results. Instead we found strong and consistent evidence for infants having a preference for patterns including immediate repetition (XXY

or XYY). It didn't make a difference if the sounds that infants heard were speech or non-speech (in this case, short segments from zebra finch songs akin to syllables in our speech).

Another study, reported in chapter 3, focused on whether the ability to learn similar patterns in the visual domain would be improved if infants knew the name for the objects forming the pattern. A set of experiments was conducted with infants between 12 and 14 months old who are just beginning to speak and who already know the meanings of some common words. Infants were exposed to XYX or XYY rules formed by sequences of three pictures of either objects for which they knew the names, objects that were familiar but for which they did not know the name, or nonsense objects they could not have seen before or have a name for. Contrary to expectations, there was no improved performance as a result of word or object familiarity in terms of rule learning ability. Infants did, however, show more attention for pictures of objects of which they knew the name during the learning phase. Some evidence for a repetition bias was also found here, but it must be further investigated, as the number of participants was small here.

A third study on infants, discussed in chapter 4, dealt with their productions and the patterns found therein. Over 60,000 utterances of children from eight different languages in their first two years of life were analyzed to better understand the types of syllable sequences infants produce during the time we probed their rule learning abilities in perception studies like the ones mentioned above. The utterances revealed that, contrary to previous theories, infant babbling does not follow a specific timeline, in which babies first produce mostly utterances containing repetitions like *bababa* followed by a period in which they produce mostly variegated, non-repetitive utterances like *badigo*. Instead, infants are producing the latter type of utterances most frequently, and from the beginning, along with the former type (called reduplicated utterances). While this is an important finding in its own right, it also provides a relevant reference point for understanding how infants may be processing the patterns we ask them to learn in the Artificial Grammar Learning studies detailed above. Previous studies have shown that infants first repeat novel syllables in their repertoire, before using them in variegated sequences, making repetition an important learning tool for babies. As such, it follows that repeated utterances are also the ones that attract attention in the AGL experiments described above.

In the last two chapters of this thesis, chapters 5 and 6, attention is turned to adult participants and their ability to learn rules. In a simple rule learning task using the same speech stimuli as in the first infant rule learning experiments, adult participants were able to generalize the grammar they were exposed to, but only under specific conditions. The way the task was presented to them determined their success rate. They could not generalize the simple rules when instructions did not specifically tell them to seek a pattern in the input. They also benefitted from participating in an active learning task that

involved feedback, rather than a passive learning task. Finally, adult participants were able to generalize from their exposure input to novel test stimuli if the test stimuli were composed of novel items only, instead of a mix of novel and familiar exposure items; this type of testing apparently allowed them to focus on one level of processing – the structural level. The results of these experiments showed that, as is the case for infants, simple rule learning is not simple for adults either. A final exploratory experiment made use of a far more complex grammar, a Lindenmayer grammar, involving hierarchical structures and therefore more akin to natural languages than the simple grammars detailed above. In two types of testing paradigms, participants showed evidence of sensitivity to this type of grammar instantiated with different types of drum sounds, indicating that humans do have complex rule learning abilities, in the general auditory domain as well, but that the experimental stimuli and task must be presented in such a way that these abilities can be manifested.

Samenvatting

Het onderzoek dat in dit proefschrift beschreven wordt maakt deel uit van een interdisciplinair project over de cognitieve oorsprong van taal. Binnen het project werd onderzocht in hoeverre menselijke vaardigheden verschillen van de vaardigheden van zangvogels, zebravinken en budgerigars, en daarnaast werd een computationeel model ontwikkeld dat de principes van het mechanisme dat gebruikt wordt om regels (in taal) te leren inzichtelijk kon maken. Dit proefschrift is een verzameling artikelen die het leren van regels, door zowel baby's als volwassenen, beschrijft vanuit verschillende perspectieven. De onderzoeksvragen zijn of de menselijke vaardigheid om taal te leren uniek is, en of mechanismen om taal te leren taalspecifiek zijn, of ook kunnen worden gebruikt op andere domeinen, zoals bijvoorbeeld het visuele domein. Dit is overwegend onderzocht door middel van experimenten waarin gebruik gemaakt werd van het "Artificial Grammar Learning" (AGL) paradigma.

In hoofdstuk 2 wordt een uitgebreide poging om de beroemde experimenten van Marcus et al. (1999) te repliceren beschreven. In die experimenten werd aangetoond dat baby's van 7 maanden oud in een korte leerfase simpele patronen zoals XYX, XXY of XYY kunnen leren in sequenties van lettergrepen (zoals le-di-le, le-le-di of le-di-di) en dat ze deze regels vervolgens konden herkennen in nieuwe sequenties van lettergrepen die niet in de leerfase van het experiment waren aangeboden. Dit werk van Marcus et al. is vaak geciteerd als het bewijs dat baby's in staat zijn om een regel te generaliseren of toe te kunnen passen op nieuwe informatie. Ondanks onze grootschalige pogingen om deze experimenten uit te breiden en exact na te bootsen weken onze resultaten af van de originele studie. In tegenstelling tot Marcus et al. (1999) vonden we consequent bewijs dat onafhankelijk van het patroon dat in de leerfase was aangeboden, baby's de voorkeur gaven aan patronen met direct herhaalde elementen (XXY of XYY). Deze voorkeur werd zowel bij sequenties van lettergrepen gevonden, als bij sequenties van korte fragmenten uit zebravinkenzang.

In een volgende studie, beschreven in hoofdstuk 3, richtten we ons op de vraag of de baby's het makkelijker vinden om patronen in het visuele domein

te herkennen, namelijk in sequenties van telkens drie plaatjes, als ze de naam van het object op het plaatje kennen. Om dit te onderzoeken hebben in diverse experimenten baby's tussen de 12 en 14 maanden oud deelgenomen, die net beginnen te praten en de betekenis van sommige veelgebruikte woorden - en de bijbehorende afbeeldingen daarvan - al kennen. Baby's kregen XYX of XYY patronen te zien bestaande uit foto's of tekeneningen van objecten waarvan ze de naam kennen, objecten die ze herkennen maar waarvan ze de naam nog niet weten of objecten die compleet nieuw of niet-bestaand zijn of waar geen naam voor is. In tegenstelling tot onze verwachtingen werd de vaardigheid om de regels te leren niet beter wanneer baby's de woorden of objecten herkenden. Wel kregen deze objecten in de leerfase de meeste aandacht van de baby's. We vonden ook weer een lichte voorkeur voor patronen met herhaalde elementen, maar deze resultaten moeten verder onderzocht worden, met een grotere groep deelnemers.

Een derde babystudie ging over de productie van spraak in de babbelfase en de patronen die daarin voorkomen. Meer dan 60.000 babbelreeksen van kinderen van 0-2 jaar oud in acht verschillende talen werden geanalyseerd om inzicht te krijgen in de sequenties van lettergrepen die baby's produceren in dezelfde periode als de bovengenoemde perceptiestudies. Deze data lieten zien dat, in tegenstelling tot eerdere bevindingen, het babbelen van baby's geen specifiek tijdlijn volgt waarbij baby's eerst met name herhaalde lettergrepen uiten zoals *bababa* gevolgd door een periode waarin ze met name gevarieerde lettergrepen uiten zoals *badigo*. Wij vonden daarentegen dat baby's naast reeksen met herhaalde lettergrepen, reeksen met gevarieerde lettergrepen van het begin af aan produceren,. Dit is niet alleen een belangrijke bevinding op zichzelf, het helpt ook te verklaren hoe baby's de patronen leren in bovengenoemde Artificial Grammar Learning studies. Eerder onderzoek liet zien dat baby's iedere nieuwe lettergreep die ze produceren eerst vaak herhalen, en daarna op kunnen nemen in een gevarieerde reeks. Herhaling lijkt dus een belangrijk leermechanisme te zijn voor baby's, en het zou dus goed kunnen dat om deze reden herhaalde lettergrepen in de bovengenoemde AGL-experimenten extra aandacht krijgen.

De laatste twee hoofdstukken van dit proefschrift richten zich op volwassen deelnemers en hun vaardigheid om artificiële grammaticale regels te leren. In hoofdstuk 5 wordt een reeks experimenten beschreven met de op het oog simpele taak van Marcus et al. om regels te leren, met dezelfde stimuli als in de baby-studies. Volwassenen bleken de grammaticale regels te kunnen generaliseren, maar enkel in specifieke omstandigheden, namelijk afhankelijk van de manier waarop de taak aan hen werd gepresenteerd. Wanneer de instructie niet specifiek vermeldde dat ze patronen in de stimuli moesten ontdekken tijdens de leerfase, konden de volwassenen de simpele regels niet generaliseren. Deelnemers waren ook beter in het ontdekken en generaliseren van patronen wanneer ze in plaats van een passieve leertaak, een actieve leertaak kregen met feedback. Volwassen deelnemers waren tot slot ook in staat om regels te generaliseren wanneer de stimuli in de testfase enkel nieuwe items bevatten, in plaats

van een mix van bekende en nieuwe items. Deze manier van testen bleek gunstig te zijn, omdat ze bij het verwerken van informatie niet afgeleid werden door welke lettergrepen ze wel of niet eerder hadden gehoord, en zich dus helemaal konden focussen op de structuur. De resultaten van deze experimenten lieten zien dat het leren van simpele regels net zoals bij baby's niet zo gemakkelijk is voor volwassenen.

In het laatste experiment, beschreven in hoofdstuk 6, maakten we gebruik van een complexe grammatica met een hiërarchische structuur, een Lindenmayer grammatica, die daarom meer lijkt op de grammatica van een natuurlijke taal dan de simpele grammatica's die we voor eerdere experimenten gebruikten. Deelnemers bleken gevoelig te zijn voor de structuur van deze grammatica, zelfs met stimuli die bestonden uit verschillende trommelgeluiden. Dit geeft aan dat mensen in staat zijn om complexe regels te leren, ook in een niet-talig domein, maar dat de experimentele stimuli en de taak op zo'n manier gepresenteerd moeten worden dat deze vaardigheid ook daadwerkelijk tot uiting kan komen.

Curriculum Vitæ

Andreea Geambaşu was born in Bucharest, Romania and grew up in San Diego, California, where she also completed her Bachelor studies (2004-2007). She double majored in Linguistics and in French Literature at UC San Diego, graduating Cum Laude. During that time she also worked as an intern in the Cognitive Development Lab of Prof. Dr. Gideon Deák (2006-2007), gaining her first experience in designing experiments for children.

After graduating from UCSD, Andreea worked as a Romanian Language Specialist in Text-to-Speech at Nuance Communications (2008) and as a Technical Editor at the academic journal *Acta Horticulturae* (2008-2009). In 2009, she enrolled in the interdisciplinary Brain and Cognitive Sciences MSc program of the University of Amsterdam. During the two year program, she conducted two research projects on language processing, the first with Prof. Dr. Annette de Groot and Dr. Peter Starreveld (2010) on visual phoneme monitoring in bilinguals (Psychology), and the second with Prof. Dr. Paul Boersma (2011) on auditory phoneme categorization in bilinguals (Phonetics). Each project yielded a research thesis. In addition, she wrote a literature thesis with Dr. Dorothy Mandell and Dr. Maartje Raijmakers on the role of social interactions between caregivers and children on phonetic and lexical acquisition.

In 2012 Andreea began her PhD with Prof. Dr. Clara C. Levelt and Prof. Dr. Carel ten Cate within the NWO project "Segments and Rules". This dissertation is a result of work in the project and has produced several submitted and published articles, as well as talks and posters.

Andreea currently works as the manager of the experimental linguistics labs at the Leiden University Centre for Linguistics.