Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

Raquel Garrido Alhama

Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC Dissertation Series DS-2017-08

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

The investigations were supported by a grant from the Netherlands Organisation for Scientific Research (NWO), Division of Humanities, to Levelt, ten Cate and Zuidema (NWO-GW 360.70.452).

Cover design by Roki.
Printed and bound by Ridderprint.

ISBN: 978-94-6299-786-8

Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 29 november 2017, te 13.00 uur

door

Raquel Garrido Alhama

geboren te Artés, Spanje

Promotiecommisie

Promotor:	Prof. Dr. C. J. ten Cate	Universiteit van Leiden
Co-promotor:	Dr. W.H. Zuidema	Universiteit van Amsterdam
Overige leden:	Prof. Dr. K. Sima'an	Universiteit van Amsterdam
	Prof. Dr. H. Honing	Universiteit van Amsterdam
	Dr. A. Alishahi	Universiteit van Tilburg
	Prof. Dr. P. Monaghan	University of Lancaster
	Dr. J. E. Rispens	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

And we know what we're knowin'
But we can't say what we've seen

Talking Heads, *Road to Nowhere*

Contents

Acknowledgments

This adventure started when Jelle, Remko, Carel and Claartje offered me a PhD position. I could not thank them enough for giving me the opportunity to work on such an interesting project and with such a skilled team of researchers, including also my fellow PhD students Andreea and Michelle. I am thankful for the open and gentle atmosphere that they created in our project meetings, and I appreciate that Carel became my promotor and helped in keeping my PhD on track.

I have been very fond of my weekly meetings with Jelle and Remko. I never stopped feeling amazed by the intelligence and originality of their ideas, but also by their passion for research, intellectual honesty and ambition to pursue interesting projects even when they are challenging. Those are some of the greatest lessons that I take from my PhD; I hope I can carry on my research in the same spirit. I am thankful to Jelle for the patience he has always had towards the chaos that accompanies my work a bit too often, for his surprising ability to find meaning in my unstructured thoughts, and for giving me so much freedom and encouragement to follow my own intuitions. I never had the chance to tell Remko how much I valued his early presence in our meetings for his sharp observations, witty humour and contagious optimism. He is very much missed.

I always talk with pride about the life I have had as a PhD student. I consider the ILLC to be a privileged environment for research, where PhD students are taken care of and well respected. I am thankful to everyone that contributes to the well functioning of the institute, and specially to Jenny and Tanja, who have been very involved whenever I needed anything. I thank Jelle and the ILLC for offering me a job as a Lecturer after my PhD, which I found professionally very enriching. I also appreciate Samira's patience and dedication as a Teaching Assistant of a course that was being designed on the fly.

But being a PhD is not an easy endeavour, and I could not escape the occasional anxiety and lack of confidence that many researchers face, nor could I avoid the dreary moods that come with the lack of sun in the Dutch weather. I am immensely grateful to Alberto, who dragged me out of bed when my energy and motivation failed me, took care of so so so many daily matters to keep me fed and warm, and put all his energy in trying to keep my mood in place. I am also thankful to Sonia, who spent hours chatting with me when I needed to cry out my mental confusion, and always reminded me to walk *pasito a pasito*. I owe huge gratitude to Miloš, who never ever denied me a single second of attention, and helped me fight my demons with rational arguments, bad jokes, Goribor, chocolate, beer, blackberries (*bam ba lam*) and disturbing photos of Tom Jones: hvala ti puno, e tako se to radi.

Luckily the ups outweigh the downs, and I could write a whole thesis just about the fun I have had during my PhD years. I consider myself very lucky for having been surrounded by extremely smart, kind and fun colleagues. I already miss my beloved office mates Sophie, Phong, Miloš, Joost, Gideon and Samira, and I am thankful to them for contributing to the nice relaxed vibe in F2.25, and for being so tolerant with my a-bit-too-loud talking and my way-too-loud typing. Miloš deserves a special mention for all the batzi papir and other silly office games, and for awesome music sessions after long hours of work; likewise, I need to thank Phong for putting up with all of that.

I am grateful to many colleagues, friends and family for great times and good laughs: to all the buddies who joined the basketball games, and to those with whom I have had intellectual discussions that were often as wild as fun; to the colleagues in the SMT group, with whom I felt like an adopted sister, and to Khalil for taking us out for drinks so many times; to Olivier, Charline and Alberto for many weekends of seriously intense boardgame marathons; to Jo and Phong for great gym sessions and even better aftergym chit chat; to Miloš for taking me repeatedly to Repetitor's gigs; to Andreea and Michelle, who were not only competent team mates but also warm friends and conference buddies; to Carmen and Marieke, with whom I have also had lots of fun in conferences (including a memorable night singing *L'Estaca* in the streets of Donostia), and whose hospitality and weekend hikes made my time in Edinburgh very dear to me; a mi família, que no sólo ayudó con una mudanza de noventa cajas y seis pisos sin ascensor, sino que ha hecho lo posible para amoldarse a mis cortas visitas y hacerme disfrutar con comilonas, chistes y mimos; to Judit and Sonia for their enthusiasm and interest on my research; to Alberto again for tons of fun, epic battles in Middle Earth, and Bob's burgers; and to my gang of lovely chickens, Phong, Jo, Miloš and Ke, for being caring friends and for sharing together our PhD journey with beers, ramen, trips, sauna, and many other forms of fun and sweetness.

This thesis is dedicated to Alberto and my enana Pequeña Lois for their unconditional support and biting.

Chapter 1

Introduction

1.1 Motivation

Human language is unarguably the most complex system of communication. Languages in the world consist of large vocabularies of words, which are combined to express complex meanings in a variety of syntactic patterns that allow for unlimited combinations. It is not surprising then that one of the questions that prominently occupies linguists is how humans can *learn* a new language.

Imagine the task from the perspective of a young infant attempting to learn her first language. Surely she is exposed to language in her environment (directed to her or not), but in order to go from the speech input to the meaning it conveys, she needs to succeed at a great number of subtasks. To begin with, speech is mostly continuous, so the learner needs to identify the pieces it is composed of; in other words, she has to *segment* the input into combinatorial meaningful units such as words and morphemes. This is in itself a complex problem, since the infant needs to identify first which of the available cues (stress patterns, prosodic contour, statistical information) are relevant for identifying word boundaries, and how to integrate them. But in order to also become productive with language, the learner needs to find out which rules govern the particular combinations of words and morphemes that she encounters. Thus, the infant must learn to *generalize* to grammatical novel productions; otherwise she could not hope to utter linguistic productions that she had not heard before. Rules are abundant in language, and appear at different levels, such as phonology, morphology and syntax. For instance, infants need to learn the constraints of their language regarding lexical categories, word order, morphological agreement, verb argument structure, etc.

Learning a language seems to be a greatly complicated endeavour, to the extent that many linguists have shared the intuition that linguistic input alone could not suffice to derive the right inductions (an argument known as *Poverty of the Stimulus*, [Chomsky, 1965, 1980]). From this perspective, it is not surprising that one of the most influential ideas on the second half of the past century was that infants must be genetically endowed with rich domain-specific linguistic knowledge (a *Universal Grammar*, Chomsky [1965, 1986], Pinker [1994], Jackendoff [2003]). Under this theory, the process of language acquisition was diminished, since the role of experience was limited to discover the values of the parameters of a greatly specified set of linguistic principles. This idea seemed to be supported by a well-known mathematical proof that shows that, in the absence of *a priori* constraints and negative evidence, linguistic input does not suffice for converging to the correct inductions [Gold, 1967]. Thus, for a long time, research on language acquisition assumed a very constrained learner, and since the role of experience was so limited, it became largely focused on the study of linguistic product rather than processing [Clark, 2009].

However, Gold's theorem is consistent with other explanations for the learnability of language. For instance, domain-general (rather than linguistic-specific) constraints on the hypothesis space could also facilitate the acquisition of correct grammatical patterns [Elman, 1998], while the absence of certain patterns in the input could be used as negative evidence [Rohde and Plaut, 1999, Regier and Gahl, 2004, MacWhinney, 2004, Clark and Lappin, 2010]; additionally, it should be taken into account that language has been shaped through cultural evolution to meet learnability pressures [Zuidema, 2003].

Syntactic theories based on this idea of strong nativism were also challenged with empirical research that employed psycholinguistic experiments and child-directed corpora (e.g. Lieven et al. [1997], MacWhinney [2000]), giving rise to alternative syntactic theories that put more emphasis on acquisition through language use [Fillmore et al., 1988, Goldberg, 1995, Croft, 2001, Tomasello, 2001]. In this dissertation, I focus on a particular class of psycholinguistic experiments that, by employing manually constructed artificial languages, led to the discovery that infants are more powerful learners than initially suspected, and thus strongly revitalized the interest for investigating the basic mechanisms behind language learning.

The experimental paradigm that I refer to is known as Artificial Language Learning (ALL; also known as Artificial Grammar Learning, or AGL). First proposed by [Reber, 1967], ALL experiments are characterized by the use of artificially constructed languages, based on a (typically small) set of "words" that have been carefully chosen. These word units are then combined, normally with the use of a pseudo-random procedure, such that they form a sample of well-formed "sentences" of the artificial language. This sample is used as the familiarization stimuli in experiments, generally played as a speech stream with controlled acoustic properties (e.g. syllables can be ensured to have the same syllable length, prosodic cues can be removed, etc.). In the test phase, subjects are normally tested with positive and negative stimuli, and their responses indicate whether they picked up on the properties that define the words or sentences in such language.

One of the questions addressed by ALL experiments is segmentation of a continuous speech stream into word units. Saffran et al. [1996a] famously showed that 8 month old infants are able to segment the words of a synthesized speech stream solely on the basis of distributional information, such as frequency of co-occurrence or transitional probabilities (and the same goes for adults [Saffran et al., 1996b]). Later, Aslin et al. [1998] showed that transitional probabilities alone sufficed for segmentation, while other studies revealed that stress patterns can also guide segmentation [Thiessen and Saffran, 2003, 2007]. This skill is triggered (both in children and adults) even when attention is hindered [Saffran et al., 1997]. Further studies investigate how adults respond to different manipulations of the artificial language, showing that longer sentences and greater vocabulary size hamper segmentation, while word repetition and Zipfian (skewed) distribution of words facilitate it [Frank et al., 2010, Kurumada et al., 2013].

Other studies have addressed the question of how language learners learn grammar-like rules and apply them to novel items that they have never encountered. One of the best known studies [Marcus et al., 1999] reports that 7-month-old infants generalize to novel items that are consistent with an identity relation between syllables (e.g. ABA or ABB patterns). Infants also generalize rules over word order at 12 month age [Gomez and Gerken, 1999]. In the case of adults, it has been shown that not all generalizations are equally easy to learn: some rules are only detected when the relevant syllables appear in edge positions [Endress et al., 2005], and repetition-based rules seem to be more accessible than ordinal rules [Endress et al., 2007].

Other studies focused on dependencies between non-adjacent items, which did not necessarily involve repetitions. For instance, Gómez [2002] find that both 18-month-olds and adults learn non-adjacent dependencies with greater success when they are exposed to input with more variability in the intermediate elements. Adults can track non-adjacent dependencies between consonants (with intervening unrelated vowel) and vowels (with an intervening unrelated consonant), but they fail to do so over syllables [Newport and Aslin, 2004]. Yet other studies have investigated how manipulations of a continuous speech stream (such as the insertion of pauses) affect segmentation and generalization based on non-adjacent dependencies [Peña et al., 2002, Onnis et al., 2005, Endress and Bonatti, 2007, Frost and Monaghan, 2016].

ALL is moreover not limited to human participants: it has been used with non-human animals, trained either with human speech or on vocalizations of their conspecifics. Results show that rats are able to segment a human speech stream based on co-occurrence frequency, although not transitional probabilities [Toro and Trobalón, 2005]; a similar result has been found for cotton-top tamarins [Hauser et al., 2001], while zebra finches benefit from the presence of pauses in the input to recognize coherent chunks of songs of their conspecifics [Spierings et al., 2015]. Likewise, the study on non-adjacent dependencies by Newport and Aslin [2004] was replicated with cotton-top tamarins [Newport et al., 2004], who exhibited certain stimulus-dependent differences, and rats [Toro and Trobalón, 2005], who showed no evidence of learning.

As for generalization, a version of the Marcus et al. [1999] study with birds shows that budgerigars can transfer XXY and XYX structures to novel items, while zebra finches only learn positional information [Spierings and ten Cate, 2016]. Much of animal research in ALL has been devoted to study whether animal species can learn syntactic rules that are beyond finite-state grammars. Fitch and Hauser [2004] show that tamarins can discriminate sequences from a finite-state language such as $(AB)^n$, but fail to do so for the context-free A^nB^n language (while humans succeed in both).

Starlings initially seemed to learn context-free grammars [Gentner et al., 2006], but it was later shown that the birds may have used alternative strategies [Van Heijningen et al., 2009]. Similarly, bengalese finches showed discrimination for sentences produced from a language with center-embedding [Abe and Watanabe, 2011], but acoustic similarity between training and test items could have guided the results [Beckers et al., 2012].

The abovementioned studies are just a small sample of some of the main results in ALL, but illustrate that these experiments are immensely helpful for characterizing many different aspects of language learning. That said, it turns out that in each of these experiments, when we look at the details, it is far from trivial to interpret these results. Precisely because the experiments address very different and concrete aspects of language, such as non-adjacent relations and segmentation, but it is not obvious how these aspects relate to each other. More generally, it is not easy to identify the properties of the cognitive mechanism(s) behind all the results. In order to progress towards a unified theory that explains these empirical data we need to complement the experimental research with a methodology that allows for testing multiple alternative hypothesis under different scenarios. I argue that the methodology we need is computational modelling.

The goal of this dissertation is to use the methodology offered by computational modelling to advance the current knowledge on the cognitive mechanisms behind ALL experiments. Thus, after providing the necessary background knowledge about this methodology, the coming chapters present different models that I have designed and experimented with, and illustrate the findings derived from these computational simulations. I now present in more detail the outline of the coming chapters.

1.2 Outline

Part of the research carried out in this dissertation involved conceiving a conceptual framework that identifies the main learning mechanisms involved in the experiments we are concerned with. Our framework proposes to characterize such process as a 3-step approach, concerning: (i) memorization of segments of the auditory input, (ii) determining the propensity to generalize, and (iii) generalization to a subset of novel input. The novelty of this conceptualization lies on linking steps (i) and (iii) –whose existence is widely agreed upon, regardless of concerns on whether they rely on the same or different computational principles– with the proposal of (ii). Thus, the 3-step approach is explained in detail in chapter 5 (when we propose and model step (ii)), but we use it nonetheless as the overarching structure of this dissertation.

Hence, the chapters in this dissertation are organized are follows:

Chapter 2 This chapter provides the necessary background to situate the work in this dissertation in the broader context of computational cognitive modelling, and it specially targets readers without much prior knowledge on the topic. It introduces what is a computational model, to then outline how computational models are used in the study of cognitive mechanisms. It then summarizes the main modelling traditions in cognitive modelling, and discusses how can we assess whether a model is a good explanation of a cognitive process.

✧ ✧ ✧

Part I: Segmentation

Chapter 3 The goal of this chapter is to propose an explanation of the mechanism responsible for segmentation. Based on my previous publications, the chapter presents a probabilistic exemplar model – the Retention&Recognition model, or R&R – that views segmentation as the result of retention and recognition of subsegments of an auditory input. Interestingly, R&R predicts a distribution of subjective frequencies of memorized subsegments that is notably skewed. I find that, thanks to this skew, the model exhibits excellent fit to data of experiments from human adults, but also from rats. The content of this chapter is based on the following publications:

Alhama, Scha, and Zuidema [2014] Rule Learning in humans and animals. *Proceedings of the International Conference on the Evolution of Language.*

Alhama, Scha, and Zuidema [2016] Memorization of sequence-segments by humans and non-human animals: the Retention-Recognition Model. *ILLC Prepublications, ILLC (University of Amsterdam), PP-2016-08.*

Alhama and Zuidema [2017b] Segmentation as Retention and Recognition: the R&R model. *Proceedings of the* 39th *Annual Conference of the Cognitive Science Society.*

Chapter 4 While the previous chapter presented and evaluated a model of segmentation based on its goodness of fit to empirical results, this chapter focuses on how does R&R compare to other models of segmentation. The goal of the chapter is not only to find which model of segmentation is a more plausible explanation of the results, but to reflect more broadly on how models of segmentation should be evaluated, based on an analysis of the consequences of assuming different evaluation criteria.

The content of this chapter is based on the following publications, although it also features new material:

Alhama, Scha, and Zuidema [2015] How should we evaluate models of segmentation in artificial language learning? *Proceedings of 13th International Conference on Cognitive Modeling.*

Alhama and Zuidema [2017b] Segmentation as Retention and Recognition: the R&R model. *Proceedings of the* 39th *Annual Conference of the Cognitive Science Society.*

Part II: Propensity to Generalize

Chapter 5 This chapter is my first approach to study under which circumstances humans generalize to novel language-like items. However, instead of focusing on the mechanism that explains *which* generalizations take place, we propose a model that quantifies the propensity of an individual to generalize to any novel item.

We therefore propose a novel conceptualization, based on a 3-step account: memorization, propensity to generalize and actual generalization. In order to quantify the propensity to generalize, we draw a parallel with the *smoothing* techniques employed in Natural Language Processing. We show that a rational model based on one such smoothing techniques (Simple Good-Turing, [Good, 1953]) offers a compelling alternative interpretation of the experimental results.

The work presented in this chapter was presented before in the following paper:

Alhama, Scha, and Zuidema [2014] Rule Learning in humans and animals. *Proceedings of the International Conference on the Evolution of Language.*

Alhama and Zuidema [2016] Generalization in Artificial Language Learning: Modelling the Propensity to Generalize. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, Association for Computational Linguistics, 2016, 64-72.*

Part III: Generalization

Chapter 6 This part of the thesis addresses the question of how humans generalize to *particular* novel items. But before delving into proposing a model for generalization, I present in this chapter a review of existing models of generalization in ALL. I identify what are the most relevant research questions that can be addressed when formalizing the problem, and I outline how the reviewed models have advanced on answering those questions, while also critically addressing what is still not convincingly solved. Based on this analysis, I put together a list of desiderata that aims to inspire future research.

The content of this chapter is based on the following manuscript:

Alhama and Zuidema [2017c]. Computational Models of Rule Learning. [*To be submitted.*]

Chapter 7 After having identified the most pressing unsolved issues on models of generalization, I advance the state of knowledge with the proposal of my own model. I argue that most models have been used as a *tabula rasa*, a simplification that comes at the cost of not successfully reproducing the empirical findings. Thus, my work focuses on investigating what is a plausible initial state for a model of generalization. I investigate two core ideas: (i) *pre-wiring* the model with minimal independently motivated biases and (ii) *pre-training* to account for relevant prior experience that could have influenced the task.

The content of this chapter is based on the following publication:

Alhama and Zuidema [2017a]. Pre-Wiring and Pre-Training: What does a neural network need to learn truly general identity rules? [*Under review.*]

✧ ✧ ✧

Chapter 8 I reflect on the main findings of this dissertation, as well as the limitations that should be tackled with future work.

❋ ✧ ❋

Chapter 2

Background

Before embarking on the modelling proposals of chapters 3, 5 and 7 it is worth reflecting first on what we want to achieve by building computational models, and what we want to avoid. I will do that by discussing three taxonomies for computational models: explanatory vs. predictive models, Marr's levels of analysis, and traditional families of cognitive models.

2.1 What is a computational model?

A computational model is a precise formulation of a system that can be simulated in a computer to study its behaviour. Computational models offer the possibility of exploring a wide range of ideas, since they can be simulated in a computer to see their consequences. Thanks to this, many different models –or different variations of one model– can be simulated to systematically compare their outputs. This does not only result in an advantage over the *quantity* of hypotheses that can be tested, but it can also result in improved *quality* of hypotheses, given that researchers are less constrained in the ideas they can formalize and simulate.

By using computational models to study a system, the system may be approached by dividing it into subcomponents, each of which may be individually studied. If such subcomponents are implemented as computational models, the interaction between them can also be simulated, and thus we can investigate how they should be integrated. This modular approach is especially relevant for the study of complex systems in which the study of the system as a whole is prohibitive.

Since the formulation of a model needs to be spelled out as a computer program, researchers are forced to be precise about the ideas embodied in the hypothesis they are testing. This can sometimes result in clarification of misunderstandings or in the identification of false dichotomies. An instance of this is presented in chapter 5, which shows how formalizing an idea as a computational model clarified some prior misconceptions and resulted in the rejection of a false dichotomy.

An additional advantage of formulating hypotheses as computational models is that the hypotheses can even involve the postulation of new concepts that are not immediately accessible to experimentation, but which can be studied (and maybe endorsed) with computer simulations. In other words, models may be used to produce sufficiency proofs for concepts that could not have been tested otherwise.

Thus, it seems clear that computational models can be of immense help in characterizing a system. Interestingly, computational models can additionally lead to unexpected conclusions. For instance, a model may predict how the system behaves in other settings, and this prediction may be empirically tested, perhaps prompting new discoveries in the field.

Finally, models may be used with different goals. Some models try to approximate a real system as much as possible, with the aim of producing very accurate quantitative predictions. These type of models are called *predictive* models, and they contrast with *explanatory* models, which trade the realism of predictive models with explanatory power. Thus, the aim of explanatory models is to achieve a better understanding of the principles governing a system. While predictive models are useful for certain applications, such as weather forecast or stock market prediction, research in cognitive science and linguistics is better served with explanatory models that shed some light in the mental processes underlying certain phenomena. These are the models that this dissertation is concerned with.

2.2 Marr's levels of analysis

My goal is to achieve a better understanding of the cognitive processes that explain language learning; concretely, through experimental results in the Artificial Language Learning paradigm. These cognitive processes have a physical realization in the neural substrate. However, the brain is a massively complex system: a great number of neurons are connected in complicated dynamic patterns that are responsible for generating all human behaviour, while also controlling the physiological regulations of the human body. In order to link the behaviour observed in the experiments with the neural activity, we need to understand which information is being represented in the brain and what is the process that eventually produces the observed behavioural output.

Given the complexity of the brain, it is useful for cognitive modellers to *abstract* away from many of the details of the wetware, and approach the problem at the level of information processing. But of course there is not a unique way to do this, so in practice computational models exhibit great variation in the level of abstraction they assume and the intended realism of the processes and representations they incorporate. I now introduce a well-known taxonomy that is useful for cognitive modellers to have some orientation on the different goals and levels of abstraction pursued by model proposals.

This taxonomy was proposed by David Marr, a neuroscientist and psychologist who investigated the human visual processing system [Marr, 1982]. According to Marr,

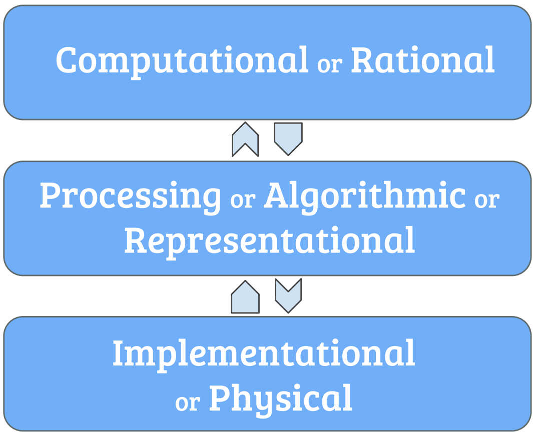

Computational or Rational
Processing or Algorithmic or Representational
Implementational or Physical

Figure 2.1: Marr's Levels of Analysis.

when studying an information system there are three explanatory levels at which we can situate models, as reflected in figure 2.1. Each of these levels is characterized by the amount of detail that is abstracted from the original system, such that the higher level (the computational level) is the more abstract, and the implementational level is the closest to the actual wetware. These levels are conceived of by Marr as complementary, with one level being a more detailed refinement of the previous one. Thus, Marr's levels of explanation offer a way to structure the problem of studying cognition by characterizing the level of simplification that the researcher may adopt.

The most abstract level is the rational or computational level. Rational models do not focus on how a task is solved, but actually aim to provide a formal description of the task itself and the strategy to solve it. As Marr puts it [Marr, 1982], the computational level is concerned with *what* is done, *why* is it done, and which is the *strategy* followed; but crucially, *how* it is done is not part of the question. For this reason, rational models often propose *optimal* solutions to a problem, that is, they identify the strategy that would offer the best performance possible given the constraints imposed by the problem.

Thus, rational models may be used as a first step to give a precise characterization of what the problem is and how it can be solved. A special class of rational models are *ideal learner* models, which investigate the problem from the perspective of an idealized observer without any limitations coming from the cognitive system (e.g. memory capacity, attention, etc.) with the aim to investigate human performance through comparison to this ideal learner baseline [Geisler, 2003].

When proposing the levels of analysis, Marr highlighted the relevance of computational level explanations:

> [...] an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied. [Marr, 1982, p.27]

So, in fact, Marr did not only offer a taxonomy to situate models, but also suggested a direction; concretely, starting at the most abstract level of analysis to eventually increase the level of detail. This approach is known as *top-down*, and it contrasts with approaches that go on the opposite direction (*bottom-up*). The arguments in favour of top-down generally state that a better understanding is achieved if starting at the more abstract functional level since the physical system is too hard to interpret, and a wrong assumption on a physical model would cause the whole system to exhibit a qualitatively different emergent behaviour (e.g. see Griffiths et al. [2010]).

In my opinion, the top-down approach can be helpful in order to constrain the space of hypotheses of what can be learnt by the system, especially for problems in which the hypothesis space is so big that we could not hope to discern the right hypothesis from emergent behaviour. However, this is unlikely to be the case in ALL, since the languages we are concerned with are very simple, and they are created with the aim to minimize regularities other than the pattern under study. Therefore, rational models of ALL generally converge very fast to learning the patterns that were originally used to design the artificial language (e.g. see an example in § 6.4), and thus they are not very revealing.

On the other hand, the processing level is well suited for modellers who aim to investigate the cognitive processes and representations underlying some phenomena –albeit without exploring their neural implementation. Models at this level of analysis are committed to postulate a mechanistic account of the steps involved in the actual cognitive process, as well as a high-level proposal of which kind of representations mediate the process. Therefore, this intermediate level of analysis is concerned with proposing a cognitively realistic account that reveals how the task is solved but without delving into details of how it translates into neural activations.

Finally, models at the implementational or physical level offer a more detailed proposal of how the processes and representations underlying a task are physically implemented in the brain. This is not to say that these models include all the physical details of its neural correlates; actually, most of these models implement a very coarse simplification of neurons and neural activation dynamics. Nonetheless, this is a very useful level of explanation to develop models that operate under constraints imposed by the general computational principles of the brain, and they can reveal unexpected emergent properties which may be later interpreted at a functional level [McClelland et al., 2010].

In this dissertation, the three models I propose belong to each of these three levels of analysis. In chapter 5 I propose a model to show that we need to account for the *propensity to generalize* in order to properly interpret the empirical data; in this case, a rational model turned out to be the more concise option to highlight the main principle behind the propensity to generalize. But in chapter 3 I opted for a model pitched at Marr's processing level, since the goal was to gain a better intuition of the process of segmentation. And finally, the model presented in chapter 7 is an implementational model, since the question it addresses (whether symbolic representations are needed to learn identity rules) required a detailed level of analysis in which the realism of those representations could play a role. Therefore, each model is pitched at the more convenient level of explanation, depending on the goals pursued in each case.

2.3 Families of Cognitive Models

I now provide an overview of the most common cognitive modelling approaches. This classification is based on well-known traditional categories of models, but of course this does not entail that all models neatly fall into one particular category. In fact, often models embrace principles from different approaches. Nevertheless, this classification is very illustrative to see of the main theoretical consequences of common modelling choices, and will be useful to situate the models proposed and reviewed in this dissertation.

2.3.1 Symbolic models

Symbolic models use discrete *symbols* to represent entities, and *rules* over symbols that represent relations. The symbols can denote observable entities —such as syllables, words or phrases— but, more generally, they can be thought of as variables or placeholders that can instantiate a certain class of entities, one at a time. For instance, if the symbol NP instantiates Noun Phrases, then NP can stand for different phrases at a given time, such as ‘my desk’, ‘Simpson’s paradox’ or ‘the infamous cat that chases the poor mouse’. Thus, the represented entities in a symbolic model may be concrete entities or abstract constructs postulated by the theory that the model embraces (which may or may not have a cognitive reality as a mental representation).

Rules are necessary to establish how symbols are related. For instance, following with our example, a rule could be S $\rightarrow$ NP VP (i.e. a Sentence is composed of a Noun Phrase and a Verb Phrase). This rule is applied over symbols, that is, the rule holds for any NP and VP, regardless of the particular content of each NP and VP. In other words, rules are *syntactic* rather than *content-sensitive*.

An important property of rules is that they are implemented as *all-or-nothing*: either they completely apply or they do not. In other words, these models do not offer graded acceptability. This can be seen as the main strength of these models; however, this also entails that these models do not exhibit graceful degradation, that is, they are not robust to small variations in the input. In order to alleviate this, some symbolic models are implemented as *probabilistic symbolic models*, in which productions are assigned a certain probability that determines their acceptability. Some examples of symbolic models are formal grammars (e.g. Context-Free Grammars), in which symbolic rules operate over terminals (words or morphemes) and non-terminals (non-observed entities that are part of linguistic theory), and define the scope of well-formed sentences. These models can be seen as implementations of generative theories of syntax [Chomsky, 1957], according to which the postulated entities are cognitively real; nevertheless, these models generally fit better at Marr’s computational level of analysis, since they concern the task (finding a proper description of the input) rather than the nature of the cognitive processes involved in finding such a description.

Other examples of symbolic models include formal approaches to semantics [Gamut, 1991]; SOAR, a model that aims to provide a unified theory of cognition [Newell, 1990]; and some components of ACT-R, a processing level model which is conceived of as a full cognitive *architecture*, i.e. a general model that implements the most basic cognitive operations [Anderson, 2014].

2.3.2 Exemplar-based Models

Exemplar models emphasize the role of memory over the role of processing. The main property of exemplar-based models is that most of the perceived input is stored, generally in a very rich representation that involves many features. Representations may vary in their complexity, so small phonetic units could be stored as well as some complete utterances. However, these representations often are restricted to observed items, that is, theoretical constructs such as NP, VP or S are often not stored in exemplar models.

Thanks to the rich representations, exemplars can be related based on some notion of *similarity*. Therefore, exemplar-based models need to come equipped with some similarity distance to relate items. For instance, a model may *recognize* a novel input as belonging to the same latent category as some other item already stored in memory if the similarity distance is small enough.

A relevant property of exemplar models is that every token exemplar is stored. This entails that we can derive the frequency of a type by counting the number of stored tokens of the same type. This is an important difference with symbolic models, in which frequency did not have any role. By considering frequencies, non clear-cut decisions can be made; for instance, frequent productions may be deemed more acceptable than infrequent ones.

All these properties make this models very robust to errors, since –contrary to symbolic models – they are content-dependent. However, one important drawback of exemplar-based models is that they do not generally handle well phenomena which appear to be very systematic. This is because most exemplar-based models do not store any form of abstract information; for instance, there may be no NP entity stored in an exemplar-based model of syntactic processing.

Regarding Marr's levels of analysis, exemplar models are a clear example of algorithmic level models, since they are based on cognitive assumptions about how information is stored and represented. Even though they may include numerically-coded statistical information, modellers assume that those have neural correlates in the form of strength of memory traces, associations and activation strength, respectively.

One of the most notable application of exemplar models to language is the work by Royal Skousen, formalized in a general exemplar-based framework called Analogical Modeling [Skousen, 1995, Skousen et al., 2002]. The main tenets of exemplar-based models are also at the core of syntactic theories that are based on the storage of linguistic *constructions* (as opposed to the storage of separate entities for lexical items and syntactic rules). This idea has crystalized in a range of proposals of grammatical formalisms that assume a rich inventory of linguistic constructions [Fillmore et al., 1988, Goldberg, 1995, Croft, 2001, Steels, 2013], and computational implementations such as Data-Oriented Parsing [Scha, 1990, Sima'an, 1999, Bod, 2006, Zuidema, 2006] (although some of these proposals include also symbolic information).

Other exemplar-based models of language investigate grammar acquisition [Batali, 1999, Borensztajn et al., 2009], stress patterns [Daelemans et al., 1994] and inflectional morphology [Keuleers and Daelemans, 2007], among others. In this dissertation, a probabilistic exemplar-based model is presented in chapter 3.

2.3.3 Bayesian Models

As mentioned before, some symbolic models relax the rigidness of all-or-none rules, by using probabilistic rules instead. These probabilities may be computed in different ways, such as derived from the frequency counts items. One particular approach to probabilistic modeling that has become very prominent in cognitive modelling is Bayesian Modelling.

Bayesian models offer a perspective for reasoning under uncertainty. Probabilities are a natural choice to model knowledge based on gradual degrees of belief, and Bayesian models provide a framework to formalize how to reason about new data based on actual knowledge or beliefs. Thus, these models conceive of learning as a problem of induction from what is known to what is not. The process of going from known to unknown facts is called *inference*.

According to this framework, when a learner faces some new data d, she tries to find an explanation for that data in terms of which process generated the data. We refer to the collection of possible hypotheses for explaining the data as H. For instance, d could be a sentence like "I saw the thief with my glasses", and H could be a set of grammatical rules that could have generated the observed linguistic input (e.g. one in which "with my glasses" is attached to "saw" and one in which it is attached to "thief"). The task of the learner is to decide which of these hypotheses (syntactic trees) is more likely to have generated the sentence. This can be formalized as

$$\arg\max_{h \in H} P(h|d) \tag{2.1}$$

That is, the goal is to find which hypothesis h has maximum probability for the observed the data. This probability is called *posterior*.

Bayesian models provide a method of inference for computing the posterior based on the observed data d and the prior knowledge of the learner. This method is based on the computation of two components: the *likelihood* and the *prior*.

The prior, which can be written as $P(h)$, refers to the biases of the learner before observing any input. In our example, the prior shows which of the two syntactic structures would the learner favour before observing the sentence d. It could be the case that both appear equally likely; in this case, the prior should be modelled as a uniform distribution that assigns equal probability to each hypothesis (and thus it will not have any effect on the posterior). If, however, one of the syntactic structures is more salient — for example because it has been observed more often in other linguistic productions— this should reflect in an assymetrical the prior.

The likelihood is the term that introduces the data in the equation. More formally, it accounts for the probability of observing the data under a certain hypotheses, $P(d|h)$. In other words, if we fix each one of the hypothesis we consider, the likelihood tells us how probable is it to observe this data for such hypothesis.

In order to compute the posterior based on the prior and the likelihood, Bayesian models make use of to Bayes rule (eq. 2.2):

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{2.2}$$

where $P(d)$ is a normalizing term, which can be computed as $\sum_{h_i \in H} P(d|h_i)P(h_i)$. This method of inference allows the learner to transform the prior knowledge into posterior knowledge after observing data.

There have been many misunderstandings regarding their level of explanation of Bayesian models, or more importantly, the cognitive realism they commit to. In most cases, Bayesian models for cognition are pitched at Marr's computational level, since they are very well suited to investigate which statistical properties in the data may be exploited by a learner, and which rational principles may be useful to solve the problem. On some other occasions, Bayesian approaches take a step further and are used for exploring the effect of assuming different representations of the input [Griffiths et al., 2010], partially entering Marr's processing level. And finally, some modellers take the stance of the so-called "Bayesian coding hypothesis", which can be summarized under the claim that populations of neurons approximate Bayesian computations [Knill and Pouget, 2004]. The existence of such a variety of approaches, in addition to many cases of unhelpful vagueness or even internal contradictions, have resulted in substantial misunderstandings regarding the intended realism of Bayesian models in general (e.g. see Bowers [2009] for an extensive discussion on this).

Even though Bayesian models are often seen as an example of symbolic models, the fact that they can apply any type of representation blurs this categorization. Since probabilities may be derived from actual frequency of occurrence, these models may be seen as sharing properties with exemplar-based models. However, it should be noted that exemplar-based models also rely heavily on rich representations of the input, which is not a common feature of Bayesian models (although can be incorporated, as proposed in Tenenbaum and Griffiths [2001]).

Bayesian Models have been extensively used to model many aspects of language, such as word segmentation [Goldwater et al., 2009], grammar learning [Bannard et al., 2009, Alishahi and Stevenson, 2008], language evolution [Kirby et al., 2007, Smith, 2009, Thompson et al., 2016] and many more.

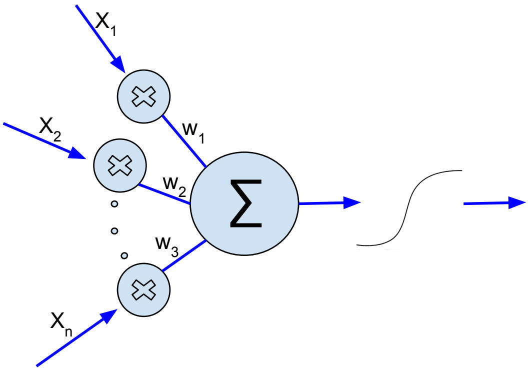


Figure 2.2: Diagram of a McCulloch-Pitts neuron, extended with a sigmoid activation function.

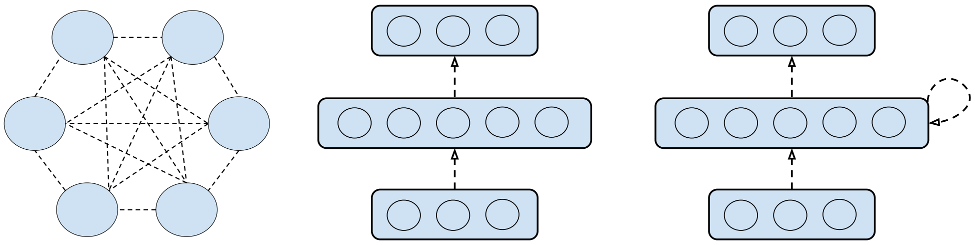
2.3.4 Connectionist models

Connectionist models, also known as neural networks, are the paradigmatic example of a model at the implementational level. This status is not without controversy, since (most) neural networks drastically simplify many implementational details; nevertheless, these models are inspired by general computational principles in the brain, so they arguably maintain the most relevant properties of the wetware.

Connectionist models consist of a network of interconnected artificial neurons (also called nodes or units) that receive and send activation signals. The simplest and most commonly used neuron model is the McCulloch-Pitts neuron [McCulloch and Pitts, 1943], which implements a coarse simplification of the functionality of a biological neuron. In this model (depicted in figure 2.2), a neuron has a set of incoming connections with an associated weight w_i. Signals $x_1, x_2, ..., x_n$ are fed into these connections, either as a result from a previous computation in a connected neuron, or as perceived input information. The input signal of each connection is scaled (multiplied) with the weight of each connection, and all the incoming signals are summed. The result of this operation is generally passed through an *activation function*; although it was a *step* function in the original formulation, it is more common to use logistic functions (such as *sigmoid* or *tanh*). After applying one of these activation functions, the model results in an upgraded non-linear node. The output of the neuron may then be passed on to other connected neurons.

A neural network consists of an interconnected set of nodes of this sort. Depending on the chosen connection topology, the architecture may take different shapes; for instance, figure 2.3 shows a fully connected network, a feedforward network, and a network that includes recurrent connections.

The information that a neural network is trained on has to be represented in vectors of activations, as reflected in figure 2.4. The most common approach is to employ *distributed* vector representations, in which each neuron participates in the representation of several items (generally reflecting a certain feature of the item, as can be seen in fig-



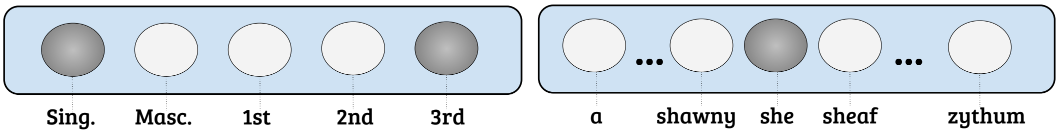
(a) Fully connected

(b) Feed-forward

(c) Recurrent

Figure 2.3: Different neural network architectures.

ure 2.4a); however, it has also been claimed that *localist* (or *one-hot*) representations, in which a neuron is only active for one single input item, may also be biologically justified (e.g. see Bowers [2009], Plaut and McClelland [2010], Quian Quiroga and Kreiman [2010] and Bowers [2010] for a discussion on the evidence supporting so-called *grandmother cells*).



(a) Distributed Representation

(b) Localist Representation

Figure 2.4: Vector representations for the word "she". In 2.4a, the word is represented based on some of its morphosyntactic features, while in 2.4b each node uniquely represents one word for the whole vocabulary.

Neural networks learn to solve tasks by gradually adjusting the weights in the connections. This is typically done based on the gradient of the error, computed with the backpropagation algorithm [Rumelhart et al., 1988]. A more biologically realistic algorithm for training the connections is Contrastive Hebbian Learning [Hebb, 1949], which is based on the Hebbian principle of "fire together, wire together" – in other words, neurons that generally fire for the same input should have their connections reinforced. In spite of the apparent differences between both algorithms, it has been shown that under some assumptions they are mathematically equivalent [Xie and Seung, 2003].

Thus, neural networks implement a domain-general learning model that extracts *associations* in the input: connections are strengthened for correlated features in the input, without any a priori defined structure. One relevant property is that abstract knowledge is not explicit in a typical neural network, although it may be implicitly learnt. Also, given that representations consist on vectors of activations, they can be seen as continuous, since they establish the coordinates of points in a multidimensional space. Additionally, as in exemplar-based models, computations are content-sensitive. For all these reasons, neural networks are generally regarded as diametrically opposed to symbolic models. This will be further discussed in chapter 6.

Neural network models have been widely applied as learning models for many aspects of language. Some examples include models of auditory word recognition [McClelland and Elman, 1986], word segmentation [French et al., 2011, French and Cottrell, 2014], and reading [Seidenberg and McClelland, 1989], but there are many more. Chapter 6 contains an extensive list of connectionist models of rule learning, and in chapter 7 I present a novel neural network model for generalization in ALL.

2.4 What constitutes a good model?

Computational models offer the possibility of investigating many different hypotheses, for which we can study their consequences by simulating them on a computer. But how can we use such simulations to assess which model constitutes a better explanation for a certain phenomenon?

First of all, it must be noted that any model that is able to reproduce a given phenomenon constitutes a *sufficiency* proof in itself. In that case, what is shown is that the theory embraced by the model is a *possible* explanation for the phenomenon. But generally many models can reproduce the same phenomenon, sometimes even based on qualitatively different principles. Therefore, even though sufficiency is not necessarily easy to achieve, it is only a minimum criterion; we need some additional method to assess which of the sufficient models is a better explanation. It should be noted though that models that fail to reproduce a phenomenon may not always be useless: sometimes models are on the right track, but an unfortunate decision on the simplification of the process may have caused an almost correct model to fail.

In the case of cognitive models, it is crucial to have *external validation*, that is, the principles embraced or ommitted by the model need to be supported with empirical evidence. But since explanatory models incorporate some degree of simplification, it is often a matter of interpretation whether the empirical data supports the proposed theory [Zuidema and de Boer, 2014].

One way to implement a form of external validation is with the use of *model parallelisation*, by comparing how multiple models explain the same phenomenon. Although this approach can be very beneficial in bringing new insights that result from the comparison between models, it does not remove the interpretative nature of the validation process. Therefore, it is still the responsibility of modellers to be critical and demanding when applying model parallelisation.

In this regard, models are often compared in terms of their output. In that case, some index of correlation between the output of the model and the empirical data is chosen in order to see which model produces an output that is closer to reality. The choice of that index should be wise: an index that is too complicated to fit may lead us to accept models that overfit the data, while an index that is too lenient or summarized (e.g. just a final average over all the responses) may not be very informative about how the models differ. This issue will come back later in this dissertation (chapter 4) with an illustration of how models embodying different principles can appear as equally good explanations for a phenomenon unless we challenge them to reproduce more fine-grained data.

There exists yet another method for evaluation. In cognitive modelling, the proposed models are a subpart of a complex system, and therefore, they should eventually interact with the rest of subcomponents of the system in a proper way. When that is the case, a model can be evaluated regarding its role in the system in which it is contained. Zuidema and de Boer [2014] refer to this as *model sequencing*, and argue that it constitutes a strong form of model validation, specially when external validation is unattainable due to lack of evidence. Chapter 5 shows an example of model evaluation based on sequencing: it demonstrates how the output of the model presented in chapter 3 has the necessary properties to make the next model in the pipeline defined by the system produce the desired output.

Finally, we should not forget that our aim is to build *explanatory* models. In this regard, it must be noted that part of the process of building explanatory models is finding a proper way to *simplify* reality; after all, the features that cause the studied behaviour may only become apparent when other less relevant features are excluded. Thus, models with too much detail obscure the properties of the system, although oversimplification could also result in incorrect predictions or in limiting the phenomena that the model can explain [McClelland, 2009]. Therefore, a cognitive model should not necessarily be considered a good model when it incorporates very precisely defined mechanisms, but rather, we should praise models which are useful *caricatures* of the real system, such that they make obvious the most relevant properties while getting rid of distracting details [Segel and Edelstein-Keshet, 2013]. After all, the goal of explanatory models of cognition is to shed new light into the underlying cognitive process.

Part I

Segmentation

Chapter 3

Segmentation as Retention and Recognition

3.1 Introduction

A crucial step in the acquisition of a spoken language is to discover what the building blocks of the speech stream are. Children perform such segmentation by paying attention to a variety of statistical and prosodic cues in the input. In this process, learning and generalization mechanisms play a role that might or might not be shared with other species, and might or might not change significantly with cognitive development. Understanding the unique ability of humans to acquire speech requires an understanding of the nature of these learning biases.

Artificial Language Learning has, over the last 20 years, become a key paradigm to study the nature of learning biases in speech segmentation and rule generalization. In experiments in this paradigm, participants are exposed to a sequence of stimuli that follow a specific pattern, designed to mimic particular aspects of speech and language, and tested on whether and under which conditions they discover the pattern. A key result in this tradition is the demonstration by Saffran et al. (1996) that children of 8 month old are sensitive to transition probabilities between syllables and can segment a speech stream based on these probabilities alone; this ability to track statistics over concrete fragments of the input is often referred to as *statistical learning*. However, these experiments do not reveal whether the underlying cognitive mechanism does operate over transitional probabilities or, instead, it performs computations of an entirely different nature but which can be described as transitional probabilities.

In order to reveal the precise underpinnings of such cognitive mechanism, it is useful to resort to computational modeling. There exist several models in the literature, which are reviewed in the next chapter. However, the models that have most successfully explained experimental results are either computational level approaches [Frank et al., 2010], which do not make any predictions about the mechanistic nature of the segmentation process, or neural network models [French et al., 2011], which do provide a realistic account of the process but are less accessible to interpretation.

In this chapter[1] I present the Retention&Recognition model (or R&R for short), a new model of segmentation in ALL that explains the memorization of subsegments of a speech stream based on the cognitive processes of retention and recognition. Pitched at Marr's processing level, my model aims to offer a simple yet intuitive explanation of the process of segmentation.

I aim for the R&R model to account for results from a variety of different experiments. I test my model on several datasets: two conditions from the Toro and Trobalón [2005] studies with rats, a variant of the baseline experiment from the Peña et al. [2002] studies, and the three internet-based experiments with human adults reported in Frank et al. [2010].

This chapter is structured as follows. I start with summarizing the relevant experimental record in § 3.2. I then present my new model, and test its fit with the experimental data (§ 3.3). I derive one important novel prediction –a skew in the frequency distributions–, which I evaluate on existing experimental data for rats (§ 3.4.1). Since I could not evaluate this prediction on existing data for humans[2], in § 3.4.2 I report results from a small, new experimental study that confirms that prediction. Finally, I evaluate my model also on 2AFC experiments with human adults (§ 3.4.3) and discuss the implications of this study (§ 3.5)

3.2 Overview of the experimental record

In this chapter I focus on three existing experiments of segmentation in ALL, and I present a variant experiment that deviates in the design of the test.

The main experiment that inspires this modelling work was presented in Peña et al. [2002]. In that study, the authors expose French-speaking adults to a stream of nonsense words, and subsequently test them to ascertain whether they can (i) segment the speech stream, and (ii) detect the underlying rules and generalize them to novel stimuli.

The "words" in these experiments are syllable triples of the form AXC, where A and C reliably predict each other while X is drawn from a set of 3 different syllables. The

[1] The work presented in this chapter is based on the following publications:

- **Alhama, Scha, and Zuidema [2014]** Rule Learning in humans and animals. *Proceedings of the International Conference on the Evolution of Language.*
- **Alhama, Scha, and Zuidema [2016]** Memorization of sequence-segments by humans and non-human animals: the Retention-Recognition Model. *ILLC Prepublications, ILLC (University of Amsterdam), PP-2016-08.*
- **Alhama and Zuidema [2017b]** Segmentation as Retention and Recognition: the R&R model. *Proceedings of the* 39^{th} *Annual Conference of the Cognitive Science Society.*

[2] Despite repeated requests, at the time I was working on this topic I could not obtain access to many of the published data. Later, researchers in the Infant Learning Lab at the University of Wisconsin - Madison kindly shared results on human experiments, but the analysis over such data is omitted in this dissertation because it is inconclusive at the moment.

words in this language are ‘puliki’, ‘puraki’ and ‘pufoki’, which are part of the same “family” of words (they share the same *A* and *C*); ‘talidu’, ‘taradu’, and ‘tafodu’, which constitute another family, and finally ‘beliga’, ‘beraga’ and ‘befoga’. In the familiarization phase, subjects heard a stream of words constructed by randomly picking these words, with the constraints that two words from the same family should not appear consecutively. In some of the experiments, subliminal pauses were inserted between subsequent words in the stream.

In the test phase of the experiments, subjects were tested on whether they showed a preference for words when contrasted to *partwords* —triples that occurred in the speech stream but that cross word-boundaries, thus having the structure *CAX* or *XCA*— . On another condition, subjects were tested on their preference for *rulewords* —triples *AYC* that conform to an attested *A_C* pattern, but with a middle syllable *Y* that did not occur in this position in the stream— vs. partwords.

In the original paper, all tests involve a forced choice task, where subjects are presented with pairs of triples and are asked which of the two was more likely to be part of the artificial language they heard in the familiarization phase. Tested after 10 minutes of exposure, the subjects show a significant preference for words over partwords, but they have no preference when they compare rulewords and partwords. If the exposure time is increased to 30 minutes, they prefer partwords to rulewords. In a third experiment, micropauses of 25 ms are added between words; now, only 2 minutes of exposure suffice for revealing a preference for rulewords. In this chapter, I focus only on the experiments that compare words to partwords, but in later chapters I explore the conditions involving rulewords (see chapter 5).

Toro and Trobalón [2005, Experiment 3A] report similar experiments with rats. The animals are exposed to a 20 minute speech stream (with or without pauses) created with the same triples used in Peña et al. [2002]. Although the rats could segment a simpler speech stream on the basis of co-occurrence frequencies, when exposed to the Peña et al. stream (without micropauses) their response rates do not differentiate between words and partwords; only with the insertion of micropauses they show a higher response rate for words. With or without micropauses, the responses to rulewords are not significantly different from the responses to partwords. Toro and Trobalón interpret this as evidence for lack of generalization —rats do generalize, but less readily than humans. But since partwords were actually present in the familiarization stream and rulewords were not, the data are consistent with a model that assumes degrees of generalization. As in the case with humans, in this chapter I focus on the segmentation experiments involving words and partwords.

I also present a variant of the baseline experiment by Peña and colleagues in which I substitute the forced choice task with an alternative test. In this set up, participants (human adults) have to answer a ‘yes/no’ question about a sequence being a word of the artificial language; each of this questions is presented together with a confidence rate about the answer. As explained in § 3.4.2, this alternative type of test reveals interesting properties in the responses per test item.

Finally, in order to evaluate the model in a bigger dataset, I make use of the experiments published in Frank et al. [2010]. In this extensive study of segmentation in human adults, the authors investigate how different properties of the stimuli can influence the performance of the participants. To do so, they manipulate the number of words in the sentences that compose the stimuli, as well as the total number of different words in the language and the amount of repetitions of each word. The results show that the length of a sentence and the number of words increase the difficulty of the task, while the amount of repetitions boosts the performance of the subjects.

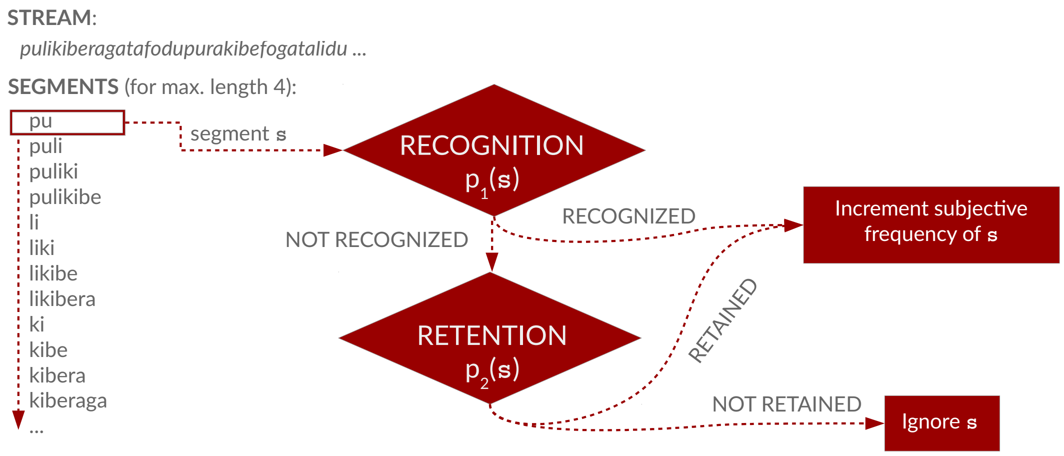


Figure 3.1: R&R:The Retention-Recognition Model

3.3 R&R: the Retention-Recognition Model

3.3.1 Model description

I present a simple processing model that describes the memorization process during the familiarization phase of the experiments. The model I propose is called the Retention-Recognition Model (R&R). It takes a sequence of syllables $X = \langle x_o, x_1, x_2, \ldots, x_m \rangle$ as input, and considers all subsequences of length $l = 1, 2, \ldots, l_{max}$ as potential segments to be memorized. Thus, the set of candidate segments is computed as shown in figure 3.2.

In the simulations reported here I assume, for computational convenience, $l_{max} = 4$. The model thus receives no specific information about the length of the words that experimenters used to create the familiarization stream (length 3 in these experiments). The model maintains a memory M, which is a set of segment types and their associated counts. The memory is initially empty ($M_0 = \emptyset$) and it changes with update steps that either *add* an entry (with count 1) or *increase* the count of an existing entry:

$$\begin{aligned}
M_{t+1} &\leftarrow \text{ADD}(M_t, \langle x_j, \ldots, x_k \rangle) \\
&\leftarrow M_t \cup \{\langle \langle x_j, \ldots, x_k \rangle, 1 \rangle\} \\
M_{t+1} &\leftarrow \text{INCREMENT}(M_t, \langle x_j, \ldots, x_k \rangle) \\
&\leftarrow M_t - \{\langle \langle x_j, \ldots, x_k \rangle, c \rangle\} \cup \{\langle \langle x_j, \ldots, x_k \rangle, c+1 \rangle\}
\end{aligned}$$

For any candidate segment $s \in S$ (with segments processed in the order they are encountered in the stream), the model checks whether it is stored in memory and, if so, what the count of that segment in memory is (its 'subjective frequency'). The model may (with a probability p_1 that increases with that count) *recognize* it (i.e., match it with a segment in memory). If it succeeds, the count is incremented with 1. If it fails to recognize the segment, the model might (with a probability p_2 that decreases with the length of the segment) still *retain* it (i.e., add it to memory with initial count of 1 if it was not stored, or increase the count by 1 as a form of 'late recognition'). In this way, the model builds a memory of segments that have different degrees of familiarity depending on their distribution in the stream. R&R's flowchart is given in figure 3.1.

The key components of the model are the equations for computing the recognition probability (p_1) and retention probability (p_2). Recognition should become more probable the more often a segment has been recognized, but decrease with the number of segment types in memory ($|M|$). Hence, I define p_1 as follows, with B and D free parameters ($0 \leqslant B, D \leqslant 1$) that can be fitted to the data:

$$p_1(s, M) = (1 - B^{\text{COUNT}(s,M)}) \cdot D^{|M|} \tag{3.1}$$

If a segment is not recognized, the model considers *retaining* it with a probability that should decrease with the length of the segment ($l(s)$), and which can be boosted if there are additional cues favoring this segment (e.g., a micropause preceding it). Hence, I define p_2 as follows, with A and C free parameters ($0 \leqslant A, C \leqslant 1$) that can be fitted to the data:

$$p_2(s, M) = A^{l(s)} \cdot C^{\tau} \tag{3.2}$$

```
Input: Stream $X = \langle x_o, x_1, x_2, \ldots, x_m \rangle$.
Output: Segments $S = \langle s_0, s_1, \ldots, s_n \rangle$.
$S \leftarrow \emptyset$
for $act = 0$ to $m$:
    for $i = 1$ to $l_{max}$:
        if $(act + i < m)$
            $S \leftarrow S \cup X[act : act + i]$
```

Figure 3.2: Pseudocode for computing candidate segments.

The A parameter thus describes how quickly the retention probability decreases with the length of a segment. The factor C^{τ} attenuates this probability unless an additional cue boosts it; here, I consider only the micropauses from Peña et al. [2002] as additional cues, and set $\tau = 0$ if there has been such a pause, and $\tau = 1$ if not. Putting everything together, the model can be described in pseudocode as in figure 3.3.

```
Input: Stream $X$, and empty memory $M_0 \leftarrow \emptyset$.
Output: Memory $M_{n+1}$.
/* Compute candidate segments: */
$S \leftarrow \langle s_0, s_1, \ldots, s_n \rangle$
/* Process each segment: */
for $i = 0$ to $n$:
    /* Compute the recognition probability: */
    $p_1 = p_1(s_i, M_i)$
    /* Compute the retention probability: */
    $p_2 = p_2(s_i, M_i)$
    /* Draw two random numbers */
    $r_1 \sim \mathcal{U}(0,1)$
    $r_2 \sim \mathcal{U}(0,1)$
    /* Recognize, retain or ignore: */
    IF $(r_1 < p_1)$
        $M_{i+1} \leftarrow$ increment$(s_i, M_i)$
    ELSE IF $(r_2 < p_2)$
        $M_{i+1} \leftarrow$ add$(s_i, M_i)$
    ELSE
        $M_{i+1} \leftarrow M_i$
```

Figure 3.3: Pseudocode describing the R&R model.

R&R is thus a simple model, but it gives a surprisingly accurate match with empirical data, as I will present in the next sections, without even taken processes such as forgetting, priming, interference and generalization into account.

3.3.2 Qualitative behaviour of the model

R&R exhibits *rich-get-richer* dynamics: as the subjective frequency of a sequence grows, the probability for this sequence to be recognized on its next occurrence in the stream also grows, and therefore its subjective frequency is likely to increase again. A sequence, however, cannot be recognized before it has been retained. The stochasticity of the retention will cause some sequences to be retained later than others, so not all

Parameters R&R	Skew Words	Skew Partwords
A=0.4 B=0.4 C=0.5 D=0.6	-0.13	2.29
A=0.5 B=0.5 C=0.5 D=0.5	0.88	1.41
A=0.6 B=0.7 C=0.7 D=0.3	-0.86	-0.37
A=0.9 B=0.4 C=0.5 D=0.6	0.68	0.0

Table 3.1: Skew for several parameter settings of R&R.

sequences will benefit equally from a high recognition probability. With this interplay between the stochasticity of retention and the (also stochastic) rich-get-richer dynamics of recognition, even sequences that are identical in terms of absolute frequency may end up with substantially different subjective frequencies.

This intuition should be reflected in the distribution of subjective frequencies. First of all, given the behaviour described above, we expect the frequency distributions to be *skewed*, since some sequences should be quickly favoured (and thus end up with high subjective frequencies) while others will be picked up later and therefore they will end up having smaller subjective frequencies. For this same reason, another prediction is that words and partwords will show *overlapping distributions*, that is, words and partwords will not be clearly differentiated given their frequencies; instead, some partwords should have higher frequencies than some words. In short, the R&R model predicts (i) a notable degree of skew and (ii) overlapping distributions of words and partwords.

I explore two different techniques to verify these predictions. In order to test for (i), I apply one metric of skew from the literature on summary statistics, concretely Pearson's second coefficient of skewness (also known as *median skew*). As for (ii), I use a qualitative approach, based on plotting the subjective frequencies of words and partwords in the same graph, in order to visualize whether the distributions overlap.

Regarding the first technique — the quantification of the degree of skew —, the median skew is defined as follows:

$$\frac{3(mean - median)}{standard\ deviation}, \tag{3.3}$$

This coefficient exploits the fact that the mean and the median are separated from each other in skewed distributions. So 0 indicates no skew, while greater values indicate greater skew, with the direction reflected in the sign.

Table 3.1 shows the median skew of a few parameter settings of the R&R model. As can be seen, the degree of skew varies substantially. This casts doubts on whether this summary statistic is a proper choice in this case, since the intuition sketched above may not be reflected in the relation between the mean and the median of the data.

As for the qualitative analysis on the distributions of words and partwords, the corresponding graph can be seen in figure 3.4, which shows the distribution of the subjective frequencies computed by the R&R model for the baseline experiment in Peña et al. [2002], for different parameter settings. As can be observed, the model presents different behaviour under different parameter settings, with most of them yielding overlapping distributions (a very large value for A is required for the distributions to be separated).

The qualitative behavior of R&R predicts therefore that the responses of subjects in the segmentation experiments will exhibit considerable degree of skew, as well as distributions that overlap. The next section analyzes whether this prediction is found in the empirical results.

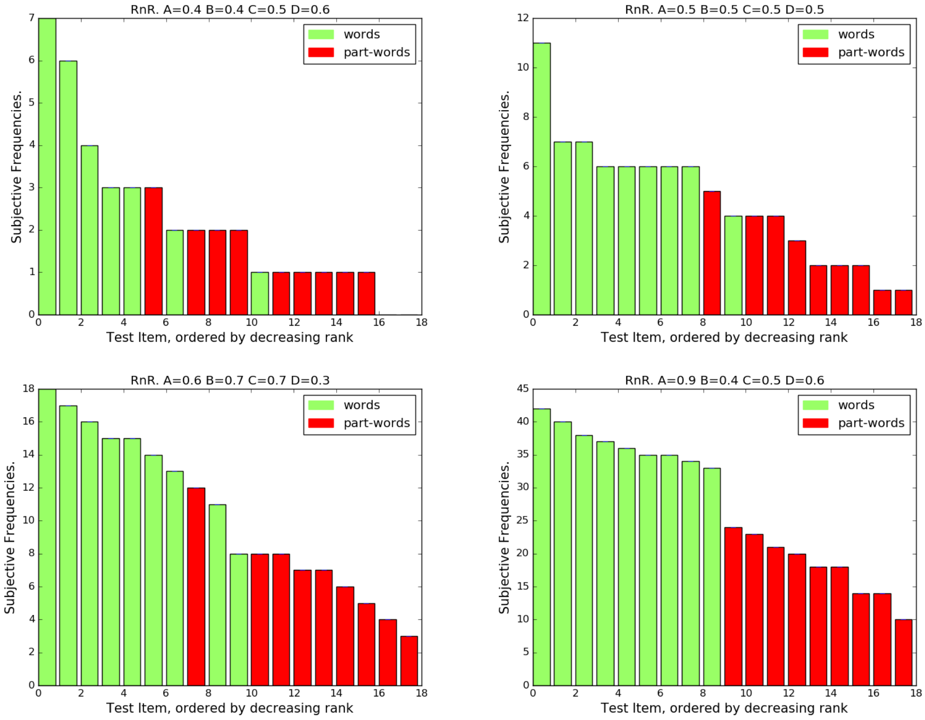


Figure 3.4: Subjective frequencies in four simulations of the R&R model, with different parameters, when familiarized with the AXC language of Peña et al. [2002]. Green bars (w) are words and red bars (pw) are partwords.

3.4 Predictions and empirical data

3.4.1 Prediction of observed skew in response distribution in rats

After having identified the main predictions of the R&R model, I now investigate whether the empirical data exhibits signatures of these predictions. In order to study whether this is the case with the experimental data from rats, I have exposed the model to the same stimuli used in Toro and Trobalón [2005], which is a stream of syllables that I created by following the description of the familiarization stream in Peña et al. [2002]. In that stream, *words* appear all with exactly the same frequency (e.g. *puliki*,

	Words	**Partwords**
No Pauses	1.01	1.07
With Pauses	1.02	1.59

Table 3.2: Median skew of the experimental data with rats.

beraga, *tafodu*, etc. appear 100 times each in the 10 minute condition). Partwords have a much lower frequency (approximately $1/2$ of the word frequency)[3].

Toro and Trobalón measure how rats respond to the test triples, based on how much they press a lever; in this way, they assess the recognition of words, partwords and rulewords. In order to test whether the responses exhibit signatures of skew, I plot the data (which the authors kindly shared) in figure 3.6. The graph shows, with small solid circles, the ordered responses of rats after familiarization to the stream without and with pauses respectively.

What can be observed in such plot is that, in line with the observations made above, R&R generates skewed distributions when presented with this familiarization stream. Such a skew has, to the best of my knowledge, not been reported yet in the analysis of experimental ALL results on adults, which all report averages over responses in a forced choice setting. Nor has it been reported in papers on experiments with prelinguistic infants and animals, which do measure responses to individual test items but all *report* averages over stimulus classes.

To evaluate how well the model can fit this data quantitatively, I make the additional assumption that the measured response rates are directly proportional to the subjective frequencies of the triples in the memories of the rats. I then search for parameter settings that produce the best fit (measured with squared error) to the average rat. I fit parameters B and D, and a value[4] that combines A and C, to the data without pauses; I then used the data with pauses solely to differentiate between the contributions of A and C. The pink lines in the graphs give the prediction of the model with the thus fitted parameters, and demonstrate a surprisingly good fit.

As explained before, I hypothesize that the observed skew in the responses (summarized in table 3.2) is explained by a skewed distribution of subjective frequencies of test items. However, an alternative explanation is that the observed skew is merely due to variability in the responses. In order to investigate this alternative explanation, I analyze how skewed would the responses be if they came from a Gaussian distribution centered on the empirical mean and scaled with the empirical standard deviation. Figure 3.5 shows a histogram of the cumulative probability for 10.000 samples. As can be

[3] The exact frequency depends on the randomization process by which *words* are sampled; in the reported simulations I have assumed that Peña et al. repeatedly play the complete sequence of words in randomized order. I also tried other processes consistent with the description they give, and obtained very similar results.

[4] As all considered segments have length 3 and there is no information to differentiate between the contributions of A and C, I estimate the value of A^3C instead. I then assume these values as given, and employ the corresponding data from the experiment with micropauses to estimate A and C.

seen, the probability of observing the empirical skew (or greater positive skew) is very small. Therefore, it seems that the observed skew is not just due to random response variability.

3.4.2 Prediction of observed skew in response distribution in humans

I have been able to confirm the prediction of skew in the response distribution of rats because I obtained access to the original data, which consisted of responses per item. But when it comes to humans, I encountered some complications: adults are typically tested in 2AFC tasks, which do not allow for a study of the distribution of responses for single items; as for infants, although the type of responses that are recorded (typically, listening times) would allow to investigate the preference for single items, the reported data consists only of averages for classes of sequences, and I could not obtain access to the original data for any of the reported studies (although see footnote 2).

For these reasons, I have run an experiment with human adults to investigate whether the skew of response distributions is consistent with that predicted by R&R. For this, I used stimuli following the structure proposed in Peña et al. [2002].

Methods

Participants

13 participants, master students of the University of Amsterdam, participated in the study as part of one of their courses.

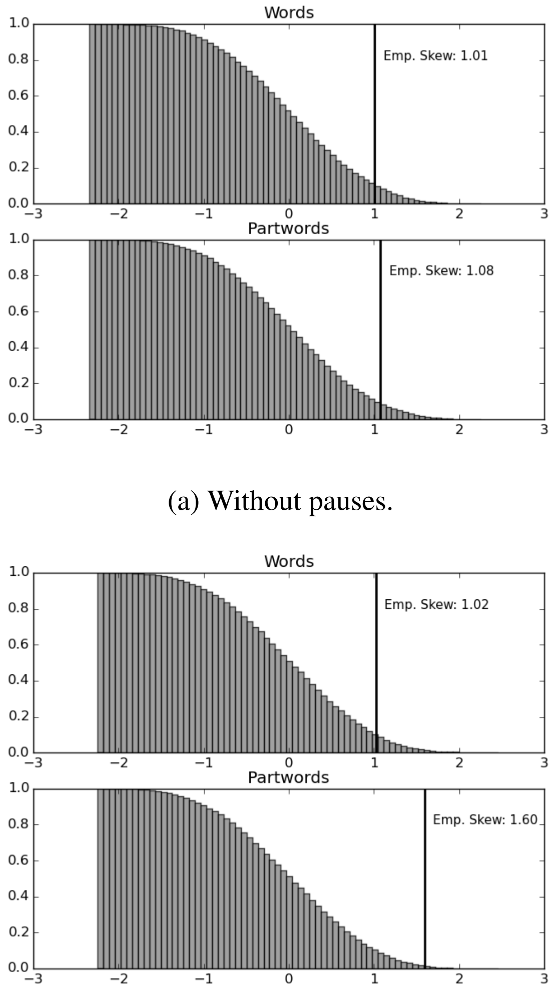
Stimuli

The stimuli consisted of an 11 minute speech stream of synthetic speech syllables generated with eSpeak. I used two conditions that only differed in the randomization of the position of a syllable in a word, and the randomization of the order of appearance of those words. For one group, the words were: *jaduki, jamaki, jataki, lidufo, limafo, litafo, sudube, sumabe, sutabe*; for the other, the words were: *jabeta, jaduta, jakita, mabefo, madufo, makifo, subeli, suduli, sukili*. Each word was presented 100 times, and their order of appearance was random with the constrain that one word cannot follow another of the same family (i.e., that starts and ends with the same syllable).

The test items consist of the nine words of the familiarization stream and nine part-words, also present in the familiarization stream, consisting of two syllables of one word and one syllable of the next, or of one syllable of one word and two syllables of the next. These eighteen items appear two times in the test set, and their order of appearance is randomized (but constant across participants), with the constraint that the same sequence does not appear consecutively.

Procedure

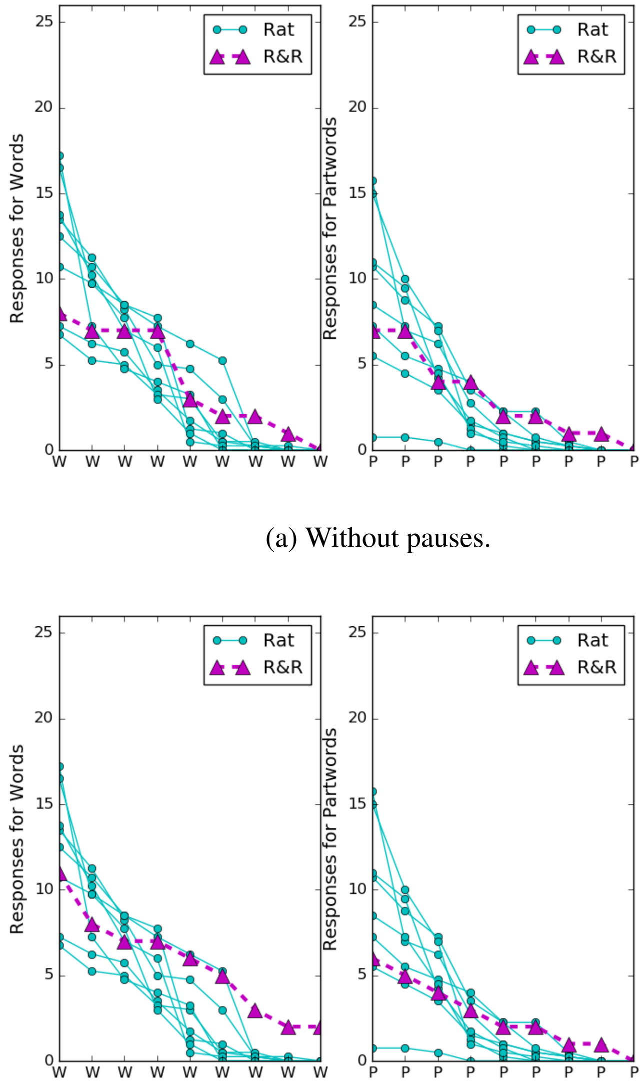
The participants were randomly assigned to one of the two conditions. The stimuli



(a) Without pauses.

(b) With pauses.

Figure 3.5: Reversed cumulative probability of the median skew coefficient, for 10.000 samples from a Gaussian distribution based on the empirical mean and standard deviation of the responses of the rats. The empirical median skew coefficient is marked with a vertical line.



(a) Without pauses.

(b) With pauses.

Figure 3.6: Responses of rats (blue) and subjective frequencies of the model (pink). W indicates words; P indicates partwords; both ordered by response frequency. Parameter setting of the model: A=0.3; B=0.92; C=0.93; D=0.94.

were presented with the use of a web form. They were instructed to listen to the whole familiarization stream, for which they would have to answer questions afterwards. In the test phase, each test item was presented acoustically, followed with the question 'Is this sequence part of the language you have heard?', to be answered with yes/no. Afterwards, they were asked to rate their confidence in the previous answer, in a scale from 1 to 7 (where 1 is minimum confidence and 7 is maximum).

Results

The average accuracy of the participants is 59.25%. This number is below that of Peña et al. [2002] (73.3%); this difference may reflect the fact that test items are presented in isolation, in contrast with the 2AFC task that Peña et al. [2002] used, where two items are presented at the same time and therefore the participant has more information (e.g., for a word that might have been accepted at chance level, the presence of its paired partword in the test can provide an extra hint for accepting the word). Nevertheless, the difference between words and partwords is significant (T-test over scale responses: t = 2.8722, df = 21.971, p-value = 0.008859).

I use the scale response of the confidence rate, multiplied with -1 if the answer to the yes/no question was negative. For each participant, I order their responses, maintaining the separation between words and partwords. Then I align the responses by their class (word or partword) and rank (position in the ordered list or responses for a particular class) and I average across participants. The assumption behind this procedure is that words are indistinguishable in terms of their frequency, but yet the most salient word for one participant need not be the most salient word for the other participant. In other words, I anonymize the particular item, while maintaining their confidence rate, rank and class.

The results are shown in figure 3.7, combining the two conditions. Thanks to analysing the responses per item, it can be observed that the data of human adults also bear out the prediction of skew.[5] Additionally, the responses given to items of the same class show a great degree of assymetry, that is, words and partwords are not clearly separated. In other words, this data meets the prediction of overlapping distributions. Thus, the results show that the skewed responses and overlapping distributions are not restricted to nonhuman animals such as rats, but are also a characteristic behavior of human adults.

3.4.3 Fitting R&R to forced choice data

We have seen that the predictions of R&R in terms of skew are visible in the experimental data of rats and humans. However, both experiments involve a small number of subjects, so I turn the attention now to the more comprehensive study by Frank et al. [2010], in order to use a sufficiently big number of datapoints to give a quantitative measure of goodness of fit of R&R. It must be noticed that, in this chapter, the evaluation applied to R&R follows the standard of the field. There are clear caveats with this evaluation procedure, but the discussion about those and the proposal for other forms

[5] This is a qualitative rather than quantitative analysis. Computing the median skew of these responses is not straightforward, given that the responses are divided in two classes (depending on the answer to the first question in the test). In the graph I have opted to distinguish those classes by casting the responses into positive or negative, but the median skew computed over such data would be affected by the fact that part of the distribution is positive and part is negative. For this reason, in this section I resort to qualitative observation of skew over the plot rather than quantitative analysis.

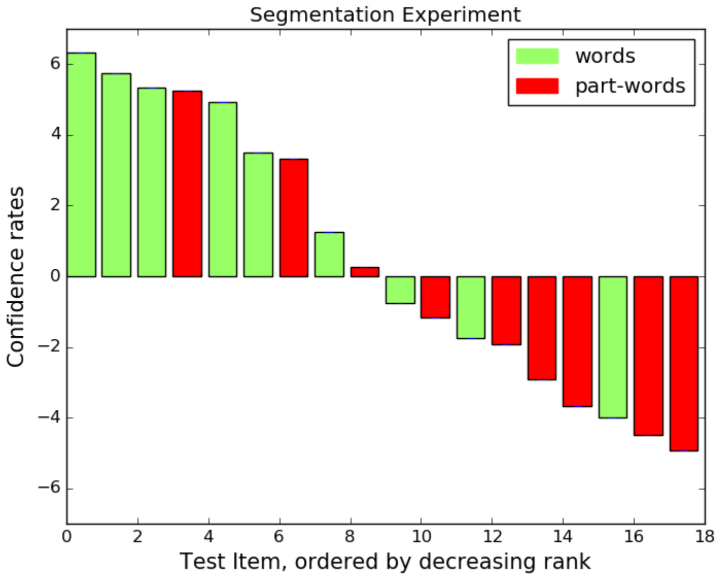


Figure 3.7: Confidence rates, averaged per ranked item.

of evaluation is left for chapter 4.

Frank et al. investigate how distributional aspects of an artificial language have an effect on the performance of human adults in segmentation. Each of their three experiments involves a range of conditions that vary in one particular dimension: (i) sentence length, (ii) amount of exposure (number of tokens) and (iii) vocabulary size (number of word types).

The stimuli consists of an auditory sequence of sentences, each of which is created from a sample of artificial (unexisting) words. The sentences are separated with a silence gap of 500 ms, while there is no acoustic nor prosodic cue indicating the separation between words within a sentence. After the participants have been exposed to a sample of sentences thus constructed, they participate in a 2-Alternative-Forced-Choice test (2AFC). The two alternatives in the test consist on one word from the artificial language (a correctly segmented sequence), and one “part-word” (a sequence resulting from incorrect segmentation).

To analyze the results, the authors average the performance (i.e. the number of correct choices) over participants. These averages are arranged to form a curve that shows the performance for different values of sentence length, amount of exposure and vocabulary size, as can be seen in the continuous line in figure 3.8. What the resulting curves show is that: (i) human adults have more difficulty in segmenting words when sentences are longer, presumably because they do not benefit from the extra cue provided by the silence gaps; (ii) when the amount of word tokens is varied, more occurrences of words facilitate the identification of such words, and (iii) the size of the vocabulary seems to cause lower performance in the experiment, with an almost-linear inverse relation.

Given that the stimuli used in this experiment contain longer pauses, the model needs some small adaptations. The design of R&R was initially inspired by the results presented in Peña et al. [2002], where the pauses in the stimuli, when present, have a length of 25ms, and this duration is supposed to be perceived by humans only subliminally. The stimuli used in Frank et al. [2010] differ significantly in the use of pauses, which have a duration of 500ms, and are used as a separation of sentences instead of words. I adapt the formula for Retention, using an exponential parameter regulating the effect of the pauses (Eq. (3.4)): [6]

$$p_2(s) = A^{length(s) \cdot \mu} \tag{3.4}$$

$$\mu = \begin{cases} \mu_{wp} & \text{after a pause} \\ \mu_{np} & \text{otherwise} \end{cases}$$

The other adaptation is the use of the Luce Rule [Luce, 1963]. Following Frank et al. [2010], I apply the Luce Rule to transform the scores produced by the R&R model (the subjective frequencies) to behavioural predictions for a 2AFC task. Given a pair of sequences s_1 and s_2 in test, the Luce Rule defines the probability of choosing s_1 as can be seen in Equation 3.5:

$$P(s_1) = \frac{SubjFreq(s_1)}{SubjFreq(s_1) + SubjFreq(s_2)} \tag{3.5}$$

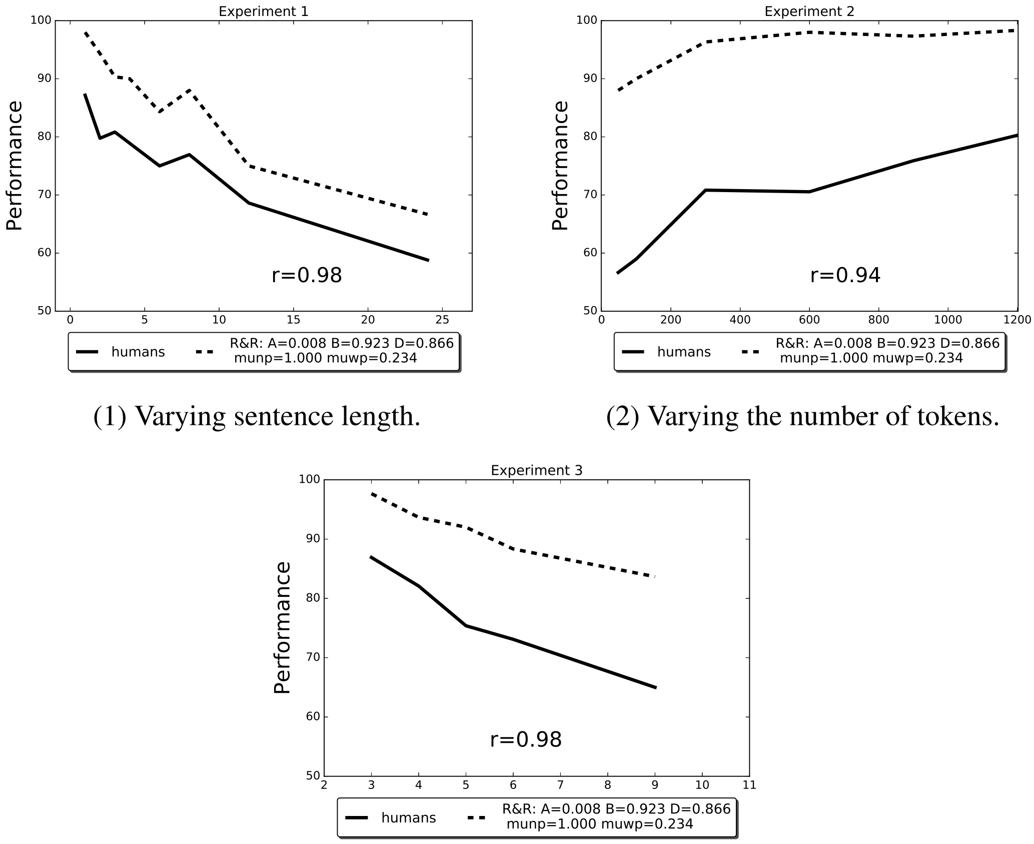
Once the scores have been transformed to probabilities, the performance of the models is computed as the mean probability of choosing the correct item, averaged over participants and test trials. These datapoints are arranged in a curve in the same way as with human participants, and the correlation in the shape of these curves — measured with Pearson's r— is taken as an indication of good fit.

The three experiments are simulated with R&R, transforming its output (the subjective frequencies) into test trials with the Luce Rule. A search is run over the parameter space, in order to find which parameters yield best correlation with human performance[7]. Clearly, optimizing the parameters on the same data on which the model is evaluated brings the risk of overfitting, but the evaluation is nevertheless carried out in this way so that the results remain comparable to other model simulations; this comparison and a discussion on better ways to evaluate are presented in chapter 4. The best results of R&R are shown in figure 3.8.

When it comes to experiment 1, one interpretation for the good fit is that R&R explicitly models the effect of the silence gaps. By increasing the length of the sentences while keeping the number of types and tokens constant, the stimuli necessarily con-

[6] As explained in footnote 4, the contributions of the two parameters A and C cannot bedistinguished, so I opted to change the formula in this way. I keep the value of the new parameter μ_{np} at 1.0 in this simulation so that the resulting model remains comparable.

[7] The only parameter that I keep fixed in the search is $\mu_{np} = 1.0$, since the interpretation of the relative importance of pauses is clearer if only one of the μ parameter is varied.



(1) Varying sentence length.

(2) Varying the number of tokens.

(3) Varying the vocabulary size.

Figure 3.8: Curve of performance for all the different conditions in the experiments in Frank et al. [2010].

sist of fewer sentences; therefore, the number of silence gaps also decreases. For this reason, the performance of R&R declines with longer sentences, since it cannot obtain the same benefit from exploiting silence gaps. This explanation can be supported by looking at the values of the μ_{wp} parameter: the best fit of the model requires a low value for this parameter ($\mu_{wp} = 0.234$)), so in the presence of a pause it substantially boosts the otherwise very small ($A^{\mu_{np}} = 0.008$) retention probability.

The second experiment was interpreted by Frank et al. as suggesting that humans may be forgetting much of what they hear, which would explain the increased performance with the number of tokens. R&R accounts for these results thanks to a probabilistic form of retention (combined with recognition that allows for the "correct" segments to be reinforced in memory); thus, the R&R model suggests that forgetting need not be incorporated in a model of segmentation, at least for the length of the stimuli used in these experiments.

Experiment 3 can be easily interpreted with R&R. The effect of increasing vocabulary size only has an effect in the distributional properties of the stream, which result in less statistically coherent partwords. Still, humans do not seem to exploit this fact; instead, their performance decreases when increasing vocabulary size. Therefore, it seems that humans have some inherent difficulty in recognizing items over a large collection of types, possibly due to interference. R&R explicitly models this phenomenon with a parameter that penalizes recognition based on the number of memorized types. In line with this intuition, the corresponding parameter value for the best fit amounts to $D = 0.86$, which substantially decreases the chance for successful recognition[8]. Therefore, in conditions of high number of types, humans have an increased difficulty in recognizing sequences, most likely originating from the process of matching the input segment to one of the many segments stored in memory.

3.5 Conclusions

Artificial Language Learning has proven to be very useful for finding out which cues are exploited when subjects are learning an unknown language. In this work I focus on one of the first problems that learners face: the identification of words in a speech stream.

With R&R, I provide a theory that considers the process of segmentation as the interaction of two cognitive mechanisms: retention and recognition. Pitched at the processing level, and with a very simple formalization, R&R offers a way to understand the pattern of experimental results that I find in the literature.

Models do not only help us reason about the cognitive processes underlying existing experiments, but also allow us to make predictions for experimental results. R&R predicts that the memorized segments of the familiarization stream should exhibit skewed and overlapping distributions of subjective frequencies; an observation that, to my knowledge, has never been reported before.

To confirm this prediction, I have revisited the experimental results of Peña et al. [2002], on human adults, and Toro and Trobalón [2005], on rats; focusing on the responses per test item: by replicating the experiment with a different test type in the former, and by providing a more fine-grained analysis in the latter. I have used qualitative and quantitative techniques to show that both datasets present skewed and overlapping distribution of responses, albeit the quantitative metric employed (the median skew) did not appear to be a very reliable measure for this prediction. Furthermore, R&R is also shown to provide a good quantitative fit to the experimental data on 2AFC responses of Frank et al. [2010].

I conclude that the R&R model constitutes a simple yet powerful characterization of the mechanisms underlying speech segmentation that shows an excellent correlation with the experimental data, and that has already allowed me to provide a new obser-

[8] Even though the values that parameter D range from 0.0 to 1.0, a value like 0.86 turns out to be relatively small. This is because the model computes $D^{|M|}$, and the number of types $|M|$ stored by R&R grows very rapidly due to the memorization of segments of any length. This entails that the whole term quickly becomes very small, so values of D that are close to 0.0 are impractical.

vation of the existing data, proving therefore to be a promising tool for revealing the properties of this basic process of language learning.

Chapter 4

How should we evaluate models of segmentation in Artificial Language Learning?

4.1 Introduction

The previous chapter presented the Retention and Recognition model, a model of segmentation in Artificial Language Learning experiments. Using existing methods for evaluating models, I showed that R&R provides a good fit to a range of empirical data. However, other models of segmentation have been proposed before, and they have also been assessed against empirical data. Thus, it is also necessary to also analyze how does R&R compare to the existing models.

The goal of this chapter is to address this issue while taking a broader perspective, that is, reflecting on general aspects of evaluation of segmentation models in ALL. As I stated in § 2.4, it is very important to challenge computational models with evaluation procedures based on empirical evidence (*external validation*, [Zuidema and de Boer, 2014]), and which are demanding enough to allow us to distinguish between different proposals.

The structure of this chapter is as follows.[1] First, I describe the most relevant models of segmentation in Artificial Language Learning, and I relate them to R&R. Second, I review how these models have been evaluated before, distinguishing between evaluation based on internal representations and evaluation over performance in 2-Alternative-Forced-Choice (2AFC) tests. The latter is further illustrated with a comparison of segmentation models, based on an extension of the study by Frank et al. [2010] –which I extend to also incorporate R&R. I reflect on what can be learnt about

[1] The content in this chapter is based on the following publications:

- **Alhama, Scha, and Zuidema [2015]** How should we evaluate models of segmentation in artificial language learning? *Proceedings of* 13^{th} *International Conference on Cognitive Modeling.*
- **Alhama and Zuidema [2017b]** Segmentation as Retention and Recognition: the R&R model. *Proceedings of the* 39^{th} *Annual Conference of the Cognitive Science Society.*

models when applying this evaluation procedure, and argue that a different type of evaluation would result in a better understanding of the differences between models. I then propose one such procedure for evaluating models, and then conclude the chapter with a call for different experiments and encouraging data sharing.

4.2 Models of Segmentation

There exist several models of segmentation in the literature; here, I focus on a representative sample, consisting of the most prominent models at each level of analysis [Marr, 1982].

At Marr's computational level of analysis we find the Bayesian Lexical Model (BLM henceforth), presented in Goldwater et al. [2006, 2009] and adapted for ALL in Frank et al. [2010]. The BLM conceptualizes the problem of segmentation from the perspective of a Bayesian model: given the input (familiarization) stream, the model attempts to reconstruct the *generative* process that generated the stream in the first place. As explained in § 2.3.3, this is done by probabilistic inference that searches through a space of possible generative hypotheses. In the case of segmentation, the hypotheses space consists of the possible segmentations, that is, each hypothesis consists of the set of words of the language that may have generated the observed stream. Thus, the 'probability of a word' refers to the probability of a sequence being a word under the current hypothesis. This is the reverse perspective from that of the R&R model. R&R is a processing model, and therefore the probabilities of sequences refer to the probability of retaining or recognizing them while processing the input.

The BLM is driven by *rich-get-richer* dynamics similar to R&R, implemented as a Dirichlet process. The main assumptions of this process are: (a) the probability of a word in the i^{th} position is proportional to the number of occurrences of this word in previous positions; (b) the relative probability for a new word type in the i^{th} position is inversely correlated with the total number of word tokens, and (c) a new word type is more probable if it is shorter.

How do these principles relate to R&R? Assumption (b) does not allow for direct comparison, since R&R is not a generative model, and therefore it does not provide a probability for new types —rather, the incorporation of new types to the memory of the model depends on the retention probability, and it is based on a preference for shorter sequences (an intuition encoded also in assumption (c) of the Bayesian model). As for assumption (a), the same principle is incorporated in the recognition process in R&R; however, in R&R the counts of the number of occurrences of a word are based on the subjective frequencies resulting from memorization, while in the BLM, these counts are based on absolute frequencies of the current hypothesis. This reflects a fundamental difference between the two approaches, which concerns their level of analysis [Marr, 1982]. The BLM is framed at Marr's computational level; therefore, since it does not incorporate any perceptual or memory constraints, it can operate over absolute frequencies (although some of the extensions in Frank et al. [2010] incorporate limitations on memory capacity, leading to somewhat hybrid models; I return to this point later). In contrast, R&R is a processing model, whose dynamics are entirely based on cognitive processes of retention and recognition, and therefore the frequency counts are a result of these mechanisms.

At the processing level, the most well known model is PARSER, a symbolic model that accounts for segmentation with basic principles of associative learning and chunking. Starting with a few primitives (typically, the syllables of the stream), PARSER incrementally builds a lexicon of segments, each of which is stored with an associated weight that has an effect on determining which segments are going to be memorized next. The size of the next segment to be perceived is determined randomly; however, the units that compose this segment will be either primitives or already-memorized segments that have a weight higher than a certain threshold. As an example, if the size of the next segment to be perceived is 2, it might be composed of two primitives (syllables), two segments (larger than the syllables) or one of each. The algorithm chooses the combination that allows the largest units, from left to right. For every new segment that is perceived, its weight in memory is incremented (or it's added with an initial weight), but the smaller units that compose it are decremented; this is meant as a process of *interference*. Additionally, at each timestep, all the units in memory have their weights decreased, as a form of *decay*.

Both PARSER and R&R are exemplar-based models that build a lexicon of segments (exemplars), and use this lexicon of already-memorized segments to decide on further segments to memorize. Each segment in the lexicon is stored together with a score that determines the impact of this segment in the next steps of the segmentation process. Thus, the models are similar in their procedure, but there are notable differences between them. One of them is the probabilistic nature of their components. For PARSER, the stochasticity is limited to the random selection of the size of the next segment to read from the stream. In contrast, R&R considers all possible subsequences of the stream (up to a maximum length), as inspired by research in Data-Oriented Parsing tradition [Scha, 1990], but is inherently probabilistic in its basic processes of retention and recognition.

There exist other differences in the procedure of these approaches. To begin with, the process of retention in R&R penalizes longest segments, on the basis that they would require more working memory. However, PARSER implements the opposite principle: whenever several segment candidates are possible, it selects those that are built of the longest units, creating in this way a bias for larger segments. As for the process of recognition, it is implicitly implemented in PARSER when it maps the next segment to be read against the units in memory. This process involves a binary threshold: only units with weight above the threshold can be recognized as components of the segment (but those below the threshold are retained). In contrast, the interaction between recognition and retention in R&R is based on a graded probabilistic choice. Finally, an important difference between the models is that R&R does not implement any form of forgetting.

On the other extreme, at Marr's implementational level, some connectionist models have been proposed in the past, based on recurrent neural networks [Cleeremans and McClelland, 1991, Servan-Schreiber et al., 1991, Christiansen et al., 1998]. More recently, an autoencoder that goes by the name of TRACX has been presented and evaluated on a range of experimental datasets [French et al., 2011, French and Cottrell, 2014]. As in all autoencoder networks, TRACX is optimized to learn a representation for the input data. Thus, the model is trained to reproduce the input in its output layer, and the error produced in the output can be interpreted as the degree of recognition of the input.

The model processes the input stream sequentially, maintaining a context window. After successful recognition of a segment, the internal representation learnt by the network is used as the context for the next segment to be presented. In this way, contiguous segments that are successfully recognized are gradually represented as *chunks*, and therefore can be recognized as a unit. This approach shares with R&R the intuition that words are consolidated in memory after repeated recognition; however, like PARSER, TRACX is a chunking model, that is, it is oriented at the integration of syllables in order to build larger fragments. In contrast, in R&R, words emerge in a process that actually penalizes larger fragments, as a consequence of consolidated memorization of statistically salient segments.

4.3 Existing evaluation procedures

Computational models need to be evaluated on some criteria to determine their validity as plausible explanations of the phenomena we aim to account for. In this section I review some of the methods of evaluation that have been used for models of segmentation in ALL.

4.3.1 Evaluation based on the internal representations

One way to evaluate the adequacy of a model is by exploring some of the properties of the representations that it builds. One example is provided in Perruchet and Vinter [1998], for the evaluation applied over PARSER. This model builds a symbolic memory of the extracted sequences, with a weight that represents the strength of their memory trace. In order to assess the quality of this built lexicon, the authors report the amount of familiarization needed to meet each of two possible criteria. The first one, called the *loose criterion*, is fulfilled when the memory contains all the words in the language with the highest weights, although it may contain also other sequences. On the other hand, the *strict criterion* states that the memory must contain all legal words with the highest weights, and if there are other memorized sequences then they must be ‘legal’ (subparts or concatenation of words, but not partwords).

The model has been evaluated based on the amount of exposure required for either criterion to be reached. For instance, Perruchet and Vinter [1998] show that PARSER meets both the loose and strict criteria with less exposure than that used for human adults in Saffran et al. [1996b], although it is not clear whether the criteria are still met after full exposure. In the case of the counterpart study with infants [Saffran et al., 1996b], more exposure than the one that infants had was required for PARSER to meet a 'looser' version of the loose criterion (concretely, that one of the two words in the test set have the higher weights in memory). Thus, a tacit assumption of this form of evaluation is that successful segmentation corresponds to having a memory in which the words have the higher weights.

The advantage of this form of evaluation is that the representations that the model builds are directly taken into account, as opposed to evaluation based on behavioural responses. However, for this to be a meaningful evaluation, the criteria used need to be independently motivated, otherwise the evaluation may be biased to favour the modellers' idiosyncratic choices. Actually, the empirical data reported later in this chapter (§ 4.5, figure 4.1) is at odds with these criteria, suggesting that this form of evaluation is not as useful as originally thought.

The evaluation procedures reported for TRACX [French et al., 2011] are also based on learnt representations, but since those are not easily observable in the model (given that they are not symbolic), the authors indirectly evaluate the representations based on the error produced by the model when attempting to recognize the represented items.

The metric (called *Proportion Better*) computes the difference between the scores for words and distractors (partwords or nonwords); in the case of TRACX, the scores are based on the recognition error of the model. This number is then normalized (see equation 4.1) and compared to the equivalent calculation for the scores of participants in the experiment, which may be based, for example, on listening times.

$$ProportionBetter = \frac{score(words) - score(partwords)}{score(words) + score(partwords)} \tag{4.1}$$

This evaluation is applied to segmentation experiments with infants [Saffran et al., 1996a, Aslin et al., 1998, French et al., 2011]. It must be noticed thought that this evaluation criterion relies on comparing the recognition error of the model, which is directly computed from the internal representations, to behavioural responses. Additionally, by averaging across test trials and participants, the whole dataset is summarized in one single number, and is therefore not very strict.

4.3.2 Evaluation based on behavioural responses

Another form of evaluation is also based on behavioural responses. Most of the experimental record on segmentation on adults involves some form of 2AFC test between target sequences (words) and distractors (generally a partword or a nonword). This kind of data is more difficult to relate to internal representations (since responses for one stimulus depend on the presentation of the alternative item in the test trial), so in order to fit this data models need to postulate an additional hypothesis –a *response model*– for linking memorized representations to the behavioral responses.

For instance, PARSER has been evaluated against 2AFC data [Perruchet and Vinter, 1998, Perruchet et al., 2004], using the following response model over the computations of PARSER: given a test trial involving a word and a partword, the response model chooses the sequence with strongest weight in the memory of PARSER; when a test trials consists of a word and a nonword (which cannot be represented in memory because they have not been encountered) then the model selects the word if it is contained in the memory of PARSER, provided that its weight is higher than 1. If the word is not in memory then the choice is random.

Thus, the response model outputs a sequence of choices over the 2AFC trials. From these, the authors compute the performance of the model, that is, the average proportion of correct choices (i.e. choices for words). This final number is compared to the performance of human participants in order to evaluate the good fit of the model.

This evaluation procedure is therefore based on this score, which summarizes the entire experiment by averaging over test trials and participants. An important drawback of this form of evaluation is that, by relying on a single datapoint, it is likely that several models exhibit a good fit. Additionally, by evaluating models in this way, we miss the chance of exploring other interesting aspects of the models, such as how the memory evolves during the experiment and what the distribution of the memorized segments is.

Some of these issues are alleviated in the evaluation procedure used in Frank et al. [2010]. As seen in chapter 3, this study is based on three 2AFC experiments on word segmentation with adults, in which the authors explore how manipulations of certain dimensions of the input affect segmentation (sentence length, amount of exposure and vocabulary size). In order to evaluate several models with these data, the authors propose to use the Luce choice rule [Luce, 1963] as a response model to link the scores produced by models (e.g. subjective frequencies) to responses in the 2AFC test. The Luce choice rule defines the probability of choosing a sequence a on a test trial involving a and b based on the relative score of a:

$$P(a) = \frac{S(a)}{S(a) + S(b)} \tag{4.2}$$

The Luce rule is applied to a range of models, and then the average probability for choosing words (that is, the probability of a correct response) is averaged over test trials and runs. This average probability is then used for comparison with human performance. However, in contrast to the evaluation procedure presented by Perruchet and colleagues, the authors do not base the evaluation only on these two numbers; instead, they compute the Pearson's r correlation for a sequence of average performances in many conditions (i.e. they compare the curves of performance for each experiment; see figure 3.8 in § 3.4.3 to recall the shape of the performance curves).

This evaluation is a great improvement over previous approaches; on the one hand because it increases the number of datapoints, and on the other hand because it allows for the study of models from a more interesting perspective. Concretely, what is evaluated is how the performance is affected by variations in the input stream, or in other words, whether increasing or decreasing the difficulty of the segmentation task (based on sentence length, exposure and vocabulary) has a similar effect on models and humans. In the next section I discuss this form of evaluation from the perspective of model selection.

4.4 Comparing alternative models against empirical data

I have discussed different forms of evaluation in the context of fitting models to empirical data. In this section, I argue that even the most adequate of the criteria discussed above can fall short when it comes to model selection. To do so, I extend the model comparison study in Frank et al. [2010] to include also the R&R model. The models simulated by Frank and colleagues include the ones previously described (BLM and PARSER, with the latter addition of TRACX, reported in French et al. [2011]), and four additional approaches: Transitional Probabilities (TP), a Bayesian version of Transitional Probabilities (Bayesian TPs), Mutual Information (MI), and a version of MI model that segments sequences that exceed a threshold both on MI and raw frequency counts (MI Clustering, Swingley [2005]). Due to their simplicity, these models are not reviewed in this chapter; it suffices to say that they all share the property of being normative models over bigrams.

Table 4.1 summarizes the goodness of fit between the models and the experimental data, based on Pearson's r correlation, as described before. As can be seen, for the parameter setting that yields better fit in the three experiments ($A = 0.008$, $B = 0.923$, $D = 0.866$, $\mu_{np} = 1.0$, $\mu_{wp} = 0.234$), the R&R model outperforms all the other models. In the previous chapter I reflected on the reasons why R&R provides a good fit to the experimental data. I now revisit and broaden that discussion to relate the outcome of R&R to that of other models.

When it comes to experiment 1, the reason for R&R outperforming other models may be that it incorporates an explicit effect of processing the pauses in the stream; concretely, in boosting the retention probability. Since the stimuli contain less pauses when sentences are longer, R&R sees its performance decreased, at a similar pace as humans. It must be noticed though that TRACX also exhibits an excellent correlation with the data, as well as two of the versions of the Bayesian Lexical Model (the original, and the one that implements uniform forgetting over types).

In the second experiment, normative models based on point estimates (those based on TP and MI) do not offer a good fit with the data, since those metrics do not benefit from the accumulation of evidence offered by the increased number of tokens (contrary to humans). Frank et al. suggest that humans may be forgetting much of what they hear, which would explain the increased performance with the number of tokens. However, the extended versions of the BLM that incorporate some form of evidence limitation (with input data restricted to a random 4% sample) or forgetting exhibit mixed results (rows 8, 9, 10, 11 on table 4.1), although the results reveal that uniform forgetting over types offers a better correlation. Still, these extensions appear unrealistic from

		Exp. 1:	**Exp. 2:**	**Exp. 3:**	
		Sentence Length	**#Tokens**	**#Types**	**Mean**
1	Transitional Probabilities	0.84	0.43	-0.99	0.09
2	Mutual Information	0.83	-0.32	-0.99	-0.16
3	MI Clustering	0.11	-0.81	0.29	-0.13
4	PARSER	0.00	0.86	0.00	0.28
5	TRACX	0.92	—	0.97	—*
6	BLM	0.94	0.89	-0.98	0.28
7	Bayesian TPs 4% data	0.82	0.92	0.96	0.90
8	BLM 4% data	0.88	0.85	0.90	0.87
9	BLM Uniform forgetting (types)	0.95	0.92	0.73	0.86
10	BLM Prop. forgetting (types)	0.88	0.87	0.88	0.87
11	BLM Uniform forgetting (tokens)	0.86	0.82	0.97	0.88
12	**R&R**	**0.98**	**0.94**	**0.98**	**0.97**

Table 4.1: Comparison of model results to human performance. The reported metric is Pearson's r. Experiment 1 is based on varying sentence length; experiment 2 on varying the number of word tokens, and experiment 3 on varying the number of word types. *Experiment 2 was not reported in French et al. [2011]. Therefore, the mean can be taken to be 0.63 (for a Pearson's r of 0.0 in experiment 2) or 0.945 (averaging only over experiments 1 and 3).

a cognitive perspective (e.g. forgetting a randomly drawn type when memory capacity is full), and additionally, the resulting models are somewhat difficult to interpret, since after incorporating memory limitations, they are different from computational level approaches. PARSER offers a more intuitive account of forgetting, with modest correlation with human data; however, this model has zero correlation in the other experiments. On the other hand, the rich-get-richer form of recognition combined with a process of retention as defined in R&R yields a better correlation than a process of recognition with forgetting.

Also on experiment 3, the R&R model exhibits the best correlation with human data, followed closely by TRACX and the Bayesian Lexical Model with uniform forgetting of tokens. Again, normative models show the opposite trend from humans (rows 1, 2, 3, 6 on table 4.1), since they do not have any memory limitations, and thus the effect of increasing vocabulary size only has an effect on the distributional properties of the stream, which result in less statistically coherent partwords. Thus, the same issues about forgetting discussed above apply to this experiment. However, with the exception of TRACX and R&R, the models that perform better in experiment 2 are not the ones that excel in experiment 3; actually, results on the BLM model suggest that uniform forgetting of tokens and 4% limitation on the input are better accounts of forgetting. Note that, as explained before, TRACX and R&R do not implement forgetting; actually, R&R explains these results based on the assumption that recognition over a large number of types is necessarily more difficult, and does so by explicitly incorporating a parameter (D) that penalizes recognition based on the number of memorized types.

But in spite of the good performance of R&R, the most relevant issue is that other models that are different also offer a good correlation under this evaluation –although never in all the experiments at the same time. Thus, we need another type of evaluation that allows for finer distinction between models.

4.5 A proposal: evaluation over response distributions

We have seen that even with the most thorough evaluation procedure that I am aware of, it is not easy to distinguish between different model proposals. On the one hand, evaluation procedures based on internal representations could be potentially useful to reveal interesting properties of the models, but it is difficult for modellers to find *external* evidence that is informative about the distributional properties of mental representations, since most of the results that could shed some light on this issue are reported as aggregated responses over participants and stimuli classes. On the other hand, model evaluation based on 2AFC responses do not allow to identify the strength of internal representations, since the observed responses for one sequence are influenced by the presence of the other sequence in the trial (e.g. an unrecognized word may be chosen because the alternative sequence may appear very unfamiliar to the subject).

I argue that, in order to evaluate and compare models, we need at least (i) exper-

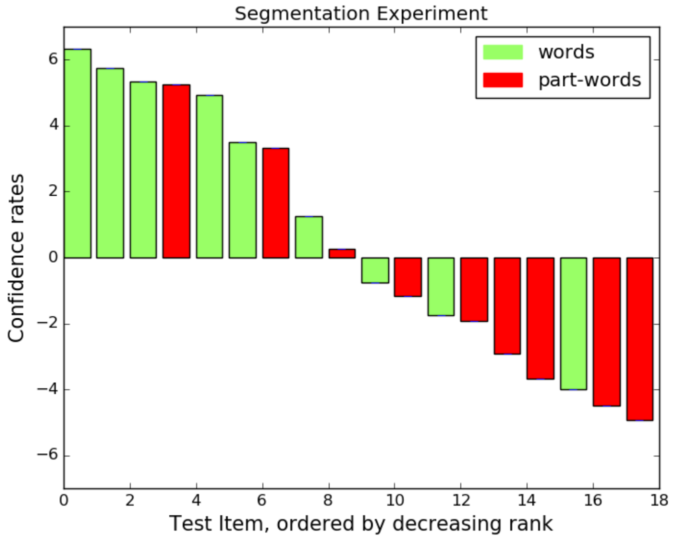


Figure 4.1: Average confidence rates for each test stimulus type, in decreasing order. Confidence rates for negative answers have negative values. (Repeated from chapter 3, Fig. 3.7, for convenience.)

imental data based on (non-binary) responses to individual stimulus (ii) a big enough number of datapoints to fit; additionally, I suggest (iii) to aggregate the responses for individual test items *anonymously*. I now elaborate on these points.

There already exist experimental paradigms that are suitable for (i), such as listening times and likert scales. However, these existing paradigms are not frequently used for ALL studies on human adults. And although listening times are very often the type of experiment employed for infants, the results reported in papers are generally averages over all participants and class of test items (e.g. mean listening times for words and nonwords [Saffran et al., 1996a]). Therefore, even though fitting a model to a single quantity is a too lenient criterion, it is often the only alternative that modellers can resort to. For this reason, in order to meet (ii), modellers need access to more fine-grained data, either in the form of raw responses (which unfortunately are rarely shared in public repositories[2]), or a different summarization of the reported data.

Related to the latter, I explore with (iii) an additional way of analyzing data, which I already advanced in § 3.4.2 (but for convenience I remind of the most relevant details here). Recall that, in such experiment, I replicated the familiarization phase of experiment 1 in Peña et al. [2002]; that is, human adult participants are exposed to a speech stream constructed with words that follow an A_iXC_i pattern. But, unlike the original experiment, the test phase in this experiment did not consist of a 2AFC ques-

[2] Although see the github repository of the Language and Cognition lab in Stanford for an example of good practice on sharing raw responses: `https://github.com/langcog`.

tionnaire; instead, each test item (word or partword) is presented in isolation, followed by two questions: first, “Is this sequence a word of the language you have heard?”, which could only be responded with a yes/no answer, and second, “How confident are you?”, which had to be answered in a likert scale that ranged from 1 (not confident) to 7 (very confident). With this type of test I can now explore how much each sequence is recognized in isolation, meeting requirement (i).

I ran two conditions of this experiment, differing only in the randomization of the words. Since I found no effect on the different conditions, I plot an aggregate of the responses, as can be seen in figure 4.1.

This graph is constructed as follows. First, for each subject and for each class of test item (words and partwords), I order the responses according to their magnitude. Second, I compute the mean response per each *anonymous* test item; that is, instead of averaging for each test stimulus (e.g. averaging the responses for sequence ‘kidada’), I average over responses to stimuli that occupy the same position on ordered responses. In this way, the preferred word for a participant is averaged together with the preferred word of other participants, and the same goes for second preferred, and so on. Third, I combine information from each class (words and partwords) by ordering the sequences based on their score while maintaining class of items identifiable (in this case it is reflected in the colour). Thus, this procedure offers an additional way to analyze data that allows monitoring effects that are not due to phonetic aspects of a stimulus, but due to stochasticity and self-reinforcement in processing. This is the type of summarization that I suggest in (iii), and it also addresses (ii) in reporting a complete distribution over test items rather than a final average for each class of stimuli.

Hence, this experiment meets all the points suggested. As shown before, conducting this experiment allowed to confirm the prediction of skew from R&R, an observation that would not have been possible otherwise. It must be noticed that this kind output is not expected from every model; for instance, I run the same analysis with PARSER and, as can be seen in figure 4.2, the distribution yielded by the model does not feature the same kind of skew: the weights for the partwords are fairly similar; additionally, words and partwords are clearly separated, rather than having overlapping distributions as in the human responses. Thus, returning to the question raised before about the effect of implementing forgetting or imperfect storage, it seems that forgetting in PARSER augments the difference between words and partwords to the extent that they become clearly separated, unlike in the human responses. In contrast, the imperfect storage provided by R&R yields a distribution of items that is closer to that observed in humans.

4.6 Conclusions

Computational models endow us with a very powerful methodology for implementing and simulating an unconstrained number of ideas that describe how cognitive mechanisms may be operating. We therefore require strict and informative evaluation proce-

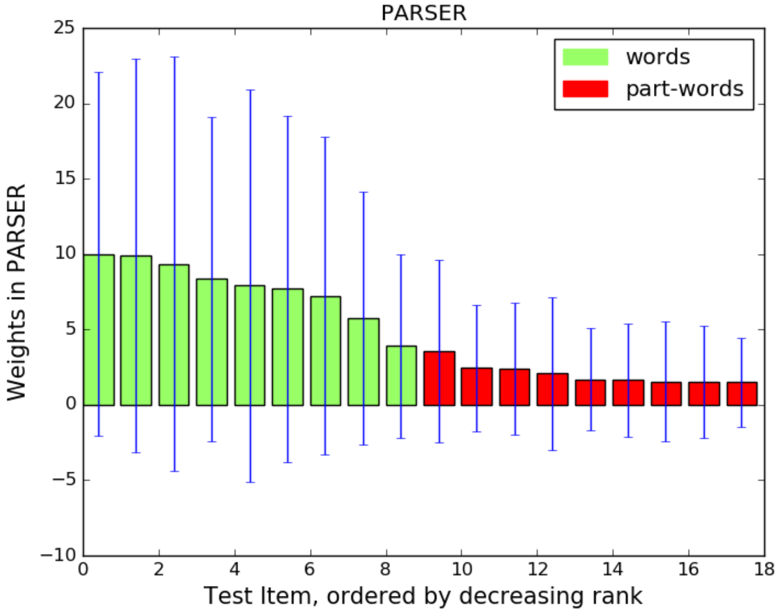


Figure 4.2: Response distribution of PARSER, over input data from experiment 1 in Peña et al. [2002].

dures to determine what constitutes a good model.

I have revised the existing evaluation procedures, and I have reached two main conclusions. First, that when evaluation is not based on quantitative fit to empirical data but rather on the researcher's intuitions on the internal states, the assumptions may be flawed. Second, that we need to rely not only on 2AFC responses but also on other type of experiments. I hope this encourages experimentalists to provide us also with experiments that allow for an interpretation of the subjective frequencies (memory traces) of individual stimulus. Likewise, modellers require access to such data after it has been published. In other fields it is common to use online open repositories to share either data or code, and it would be very useful if that became the standard in ALL as well.

Finally, I have suggested a complementary form of analysis over experimental data. In the ALL literature, responses are typically tested in order to see if subjects respond differently to classes of stimuli (such as words vs. partwords). While this analysis is necessary to assess hypotheses such as whether subjects can identify words in a speech stream, it overshadows other interesting behaviours. Thus, I have proposed to analyse responses based on aggregating the data *anonymously*, with the intuition that subjects may have idiosyncratic preferences over stimuli. This prompted the discovery that distributions of memorized sequences exhibit overlap between classes of stimuli, and are more skewed than initially expected. The procedure reported here constitutes one proposal that has already shown to be useful to show that the process of segmentation exhibits some form of self-reinforcement (*rich-get-richer*) dynamics, but other methods should be explored in order to challenge computational models with strict evaluation criteria.

Part II

Propensity to Generalize

Chapter 5

Modelling the Propensity to Generalize

In the previous chapters I have proposed a model for segmentation in ALL, which I have contrasted with other existing models. But segmentation is not the only task that language learners need to master; as advanced in § 1.2, it is useful to conceptualize (some of) the mechanisms involved in language learning as a three-step approach that starts with segmentation and culminates in generalization to unseen productions. In this chapter, I motivate the need for accounting for the second step: the propensity to generalize.

5.1 Introduction

First of all, let's recapitulate on the main findings in the field. In the last 20 years, ALL experiments have become increasingly popular for the study of the basic mechanisms that operate when subjects are exposed to language-like stimuli. Thanks to these experiments, we know that 8 month old infants can segment a speech stream by extracting statistical information of the input, such as the transitional probabilities between adjacent syllables [Saffran et al., 1996a, Aslin et al., 1998]. This ability also seems to be present in human adults Saffran et al. [1996b], and to some extent in nonhuman animals like cotton-top tamarins [Hauser et al., 2001] and rats [Toro and Trobalón, 2005].

Even though this statistical mechanism is well attested for segmentation, it has been claimed that it does not suffice for generalization to novel stimuli or *rule learning*[1]. Ignited by a study by Marcus et al. [1999], which postulated the existence of an additional *rule-based* mechanism for generalization, a vigorous debate emerged around the question of whether the evidence from ALL-experiments supports the existence of a specialized mechanism for generalization [Peña et al., 2002, Onnis et al., 2005, Endress and Bonatti, 2007, Frost and Monaghan, 2016, Endress and Bonatti,

[1] I prefer the term 'generalization' because 'rule-learning' can be confused with a particular theory of generalization that claims that the mental structures used in the generalization process have the form of algebraic rules.

2016], echoing earlier debates about the supposed dichotomy between rules and statistics [Chomsky, 1957, McClelland and Elman, 1986, Pinker and Prince, 1988, Pereira, 2000].

In this chapter[2] I argue that the dichotomy between rules and statistics is unhelpful; as an alternative, I propose a different conceptualization of the steps involved in generalization in ALL. In the following sections, I will first review some of the experimental data that has been interpreted as evidence for an additional generalization mechanism [Peña et al., 2002, Endress and Bonatti, 2007, Frost and Monaghan, 2016]. I then reframe the interpretation of those results with the three-step approach, a proposal of the main steps that are required for generalization, involving: (i) memorization of segments of the input, (ii) computation of the probability for unseen sequences, and (iii) distribution of this probability among particular unseen sequences. I model the first step with the *Retention&Recognition* model. I propose that a rational characterization of the second step can be accomplished with the use of *smoothing* techniques (which I further demonstrate with the use of the Simple Good-Turing method, [Good, 1953, Gale and Sampson, 1995]. I then argue that the modelling results shown in these two steps already account for the key aspects of the experimental data; and importantly, it removes the need to postulate an additional, separate generalization mechanism.

5.2 Experimental Record

Some of the experiments that I review here have been explained before in previous chapters of this dissertation. I nevertheless report all the details that are relevant to this chapter –in spite of the overlap–, for convenience to the reader.

Peña et al. [2002] conduct a series of Artificial Language Learning experiments in which French-speaking adults are familiarized to a synthesized speech stream consisting of a sequence of artificial *words*. Each of these words contains three syllables A_iXC_i such that the A_i syllable always co-occurs with the C_i syllable (as indicated by the subindex i). This forms a consistent pattern (a "rule") consisting in a non-adjacent dependency between A_i and C_i, with a middle syllable X that varies. The order of the words in the stream is randomized, with the constraint that words do not appear consecutively if they either: (i) belong to the same "family" (i.e., they have the same A_i and C_i syllables), or (ii) they have the same middle syllable X.

After the familiarization phase, the participants respond a two-alternative forced

[2] The work presented in this chapter was presented before in the following paper:

- **Alhama, Scha, and Zuidema [2014]** Rule Learning in humans and animals. *Proceedings of the International Conference on the Evolution of Language.*
- **Alhama and Zuidema [2016]** Generalization in Artificial Language Learning: Modelling the Propensity to Generalize. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, Association for Computational Linguistics, 2016, 64-72.*

stream	pulikiberagatafodupuraki..
words A_iXC_i	puliki, beraga, tafodu, ...
part-words C_jA_iX, XC_iA_j	kibera, ragata, gatafo, ...
rule-words A_iYC_i	pubeki, beduga, takidu, ...

Table 5.1: Summary of the stimuli used in the depicted experiments.

choice test. The two-alternatives involve a word vs. a *part-word*, or a word vs. a *rule-word*, and the participants are asked to judge which item seemed to them more like a word of the imaginary language they had listened to. A part-word is an ill-segmented sequence of the form XC_iA_j or C_iA_jX; a choice for a part-word over a word is assumed to indicate that the word was not correctly extracted from the stream. A rule-word is a rule-obeying sequence that involves a "novel" middle syllable Y (meaning that Y did not appear in the stream as an X, although it did appear as an A or C). Rule-words are therefore a particular generalization from words. Table 5.1 shows examples of these type of test items.

In their baseline experiment, the authors expose the participants to a 10 minute stream of A_iXC_i words. In the subsequent test phase, the subjects show a significant preference for words over part-words, proving that the words could be segmented out of the familiarization stream. In a second experiment the same setup is used, with the exception that the test now involves a choice between a part-word and a rule-word. The subjects' responses in this experiment do not show a significant preference for either part-words or rule-words, suggesting that participants do not generalize to novel grammatical sequences. However, when the authors, in a third experiment, insert micropauses of 25ms between the words, the participants do show a preference for rule-words over part-words. A shorter familiarization (2 minutes) containing micropauses also results in a preference for rule-words; in contrast, a longer familiarization (30 minutes) without the micropauses results in a preference for part-words. In short, the presence of micropauses seems to facilitate generalization to rule-words, while the amount of exposure time correlates negatively with this capacity.

Endress and Bonatti [2007] report a range of experiments with the same familiarization procedure used by Peña et al. However, their test for generalization is based on *class-words*: unseen sequences that start with a syllable of class "A" and end with a syllable of class "C", but with A and C not appearing in the same triplet in the familiarization (and therefore not forming a nonadjacent dependency between the particular syllables).

From the extensive list of experiments conducted by the authors, I will refer only to those that test the preference between words and class-words, for different amounts of exposure time. The results for those experiments (illustrated in figure 5.1) also show that the preference for generalized sequences decreases with exposure time. For short exposures (2 and 10 minutes) there is a significant preference for class-words (as can be seen by the high proportion of choices for generalized sequences); when the exposure time is increased to 30 minutes, there is no preference for either type of sequence (the generalization percentage is around chance level), and in an exposure of 60 minutes the preference reverses to part-words (as can be seen by the small proportion of choices for generalized sequences).

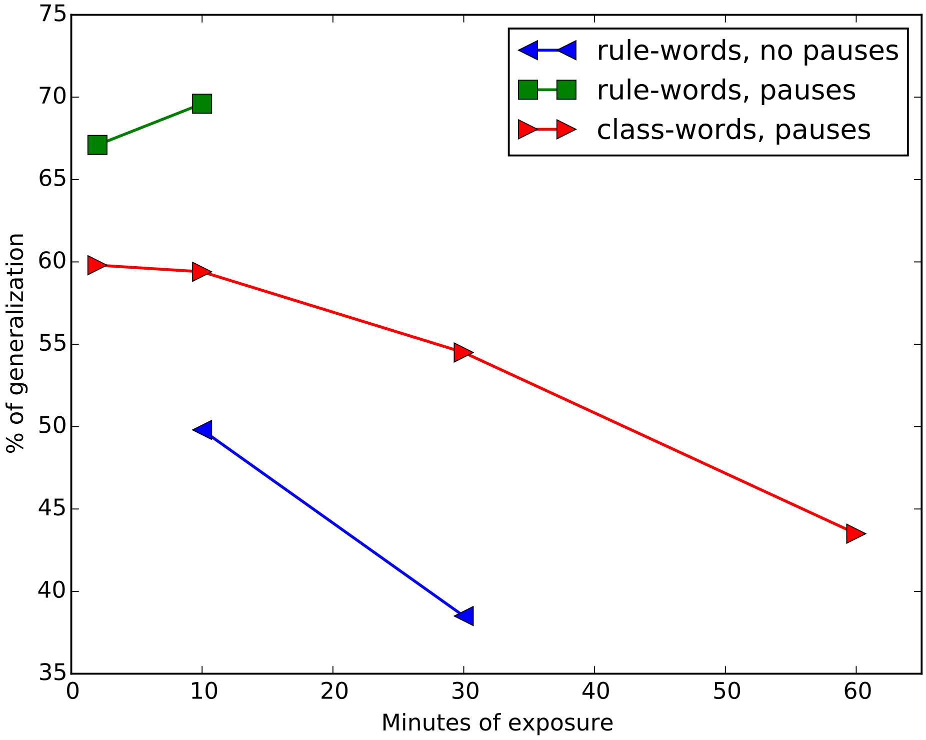


Figure 5.1: Percentage of choices for rule-words and class-words, in the experiments reported in Peña et al. [2002] and Endress and Bonatti [2007], for different exposure times to the familiarization stream.

Finally, Frost and Monaghan [2016] show that micropauses are not essential for rule-like generalization to occur. Rather, the degree of generalization depends on the type of test sequences. The authors notice that the middle syllables used in rule-words might actually discourage generalization, since those syllables appear in a different position in the stream. Therefore, they test their participants with *rule*-words*: sequences of the form A_iZC_i, where A_i and C_i co-occur in the stream, and Z does not appear. After a 10 minute exposure without pauses, participants show a clear preference for the rule*-words over part-words of the form ZC_iA_j or C_iA_jZ.

The pattern of results is complex, but we can identify the following key findings: (i) generalization for a stream without pauses is only manifested for rule*-words, but not for rule-words or class-words; (ii) the preference for rule-words and class-words is boosted if micropauses are present; (iii) increasing the amount of exposure time correlates negatively with generalization to rule-words and class-words (with differences depending on the type of generalization and the presence of micropauses, as can be seen in figure 5.1). This last phenomenon, which I call the *time effect*, is precisely the aspect I want to explain in this work. (Note, in figure 5.1, that in the case of rule-words and pauses, the amount of generalization increases a tiny bit with exposure time, contrary to the time effect. I could not test whether this is a significant difference, for lack of access to the data. Endress&Bonatti, however, provided convincing statistical analysis supporting a significant inverse correlation between exposure time and generalization to class-words).

5.3 Understanding the generalization mechanism: a three-step approach

Peña et al. interpret their findings as support for the theory that there are at least two mechanisms, which get activated in the human brain based on different cues in the input. Endress and Bonatti adopt that conclusion (and name it the *More-than-One-Mechanism* hypothesis, or *MoM*), and moreover claim that this additional mechanism cannot be based on statistical computations. The authors predict that statistical learning would benefit from increasing the amount of exposure:

> *"If participants compute the generalizations by a single associationist mechanism, then they should benefit from an increase in exposure, because longer experience should strengthen the representations built by associative learning (whatever these representations may be).*" [Endress and Bonatti, 2007]

I think this argument is based on a wrong premise: stronger representations do not necessarily entail greater generalization. On the contrary, I argue that even very basic models of statistical smoothing make the opposite prediction. To demonstrate this in a model that can be compared to empirical data, I propose to think about the process of generalization in ALL as involving the following steps (illustrated also in figure 5.2):

(i) **Memorization:** Build up a memory store of segments with frequency information (i.e., compute subjective frequencies).

(ii) **Quantification of the propensity to generalize:** Depending on the frequency information from (i), decide how likely are other unseen types.

(iii) **Distribution of probability over possible generalizations:** Distribute the probability for unseen types computed in (ii), assigning a probability to each generalized sequence.

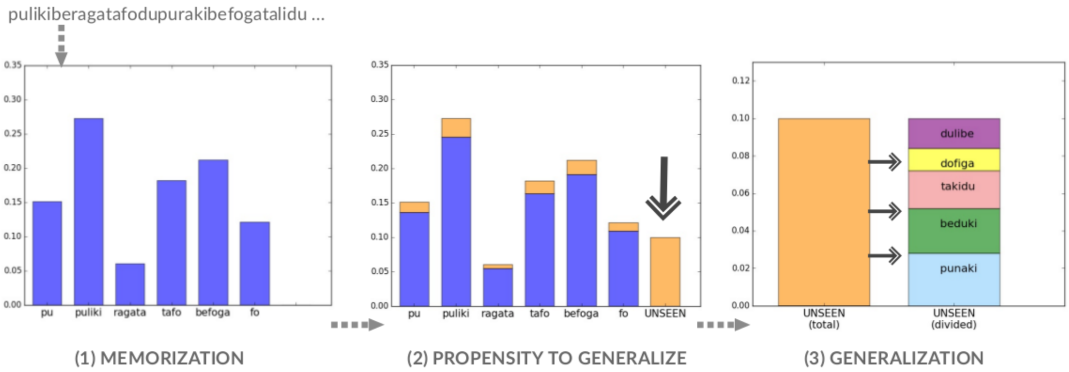


Figure 5.2: Three step approach to generalization: (1) memorization of segments, (2) compute probability of new items, and (3) distribute probability between possible new items.

Crucially, I believe that step (ii) has been neglected in ALL models of generalization. This step accounts for the fact that generalization is not only based on the particular structure underlying the stimuli, but also depends on the statistical properties of the input.

At this point, we can already call into question the MoM hypothesis: as stated in this hypothesis, more exposure time does entail better representation of the stimuli (as would be reflected in step (i)); however, contrary to what is stated in the MoM hypothesis, the impact of exposure time on generalization depends on the model used for step (ii). Next, I show that a cognitive model of step (i) and a rational statistical model of step (ii) already account for the *time effect*.

5.4 Memorization of segments

For step (i) we could choose a segmentation model from the range of models that reviewed in § 4.2, but since R&R was the most favoured in the analysis, I opt for applying it in these simulations.

Among the properties of R&R, one that is particularly relevant for this study is the *skew* that can be observed in the subjective frequencies computed by the model (see § 3.3.2). This feature is in consonance with the empirical data. Here, I show that this property can also be validated in a different way: when R&R is part of a pipeline of models (like the three-step approach), the skew turns out to be a necessary property for the success of the next model in the sequence. I come back to this point in section 5.6.

I analyze the effect of the different conditions (exposure time and presence of pauses) in the memorization of segments computed with R&R. Figure 5.3 shows the presence of test items (the nine words and nine possible part-words) in the memory of

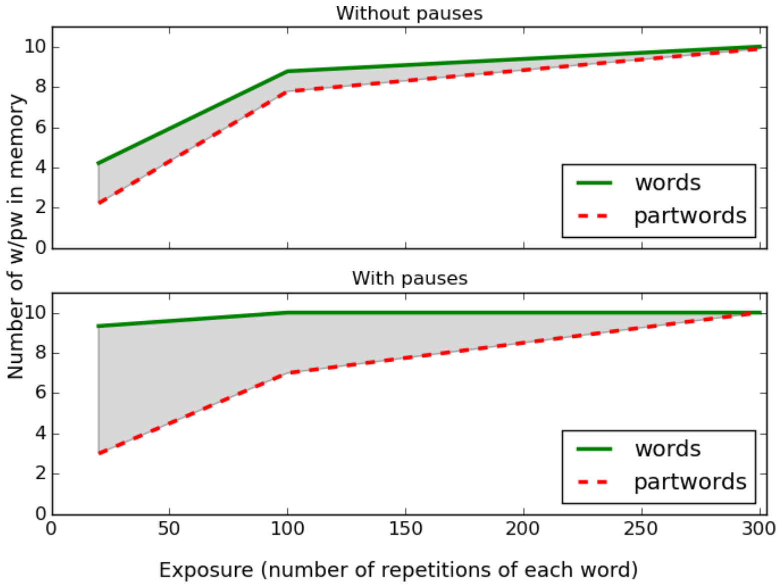


Figure 5.3: Average number of memorized words and part-words after familiarization with the stimuli in Peña et al., for 10 runs of the R&R model with an arbitrary parameter setting (A=0.5 B=0.5 C=0.2 D=0.5).

R&R after different exposure times (average out of ten runs of the model). As can be seen, the subjective frequencies of part-words increase over time, and thus, the difference between words and part-words decreases as the exposure increases.

The graph also shows that, when the micropauses are present, words are readily identified after much less exposure, yielding clearer differences in subjective frequencies between words and part-words.

The results of these simulations are consistent with the experimental results: the choice for words (or sequences generalized from words) against part-words should benefit from shorter exposures and from the presence of the micropauses. Now, given the subjective frequencies, how can we compute the propensity to generalize?

5.5 Quantifying the propensity to generalize

5.5.1 The Simple Good-Turing method

In probabilistic modelling, generalization must necessarily involve shifting probability mass from attested to unattested events. This is a well known problem in Natural Language Processing, and the techniques to deal with it are known as *smoothing*. Here, I explore the use of the Simple Good-Turing [Gale and Sampson, 1995] smoothing method as a computational level characterization of the propensity to generalize.

Simple Good-Turing (SGT), a computationally efficient implementation of the Good-Turing method [Good, 1953], is a technique to estimate the frequency of unseen types, based on the frequency of already observed types. The intuition behind this method is as follows. Imagine that a biologist goes to the forest, notebook in hand, to write down the number of animals of different species that she encounters (e.g. 20 snails, 10 sparrows, 2 squirrels and 1 snake). This data is a random sample of the animal population in the forest, and we use these frequency counts to estimate the probability that the next animal to be found belongs to one of these species (e.g. the chance that the next animal is a sparrow would be $10/33 = 0.4$). However, under this probabilistic method, the probability of observing a previously unseen animal (e.g. a rabbit) would be zero ($0/33 = 0$). What Good —acknowledging inspiration from unpublished work by Alan Turing— argues is that this is not a good estimation; instead, we should shift some probability mass from the distribution of seen species to estimate the probability of witnessing an animal belonging to an unseen species.

In order to provide a good estimation for unseen events, Good (and Turing) propose the following method: first, we take the subjective frequencies r computed by R&R and, for each of them, we compute the frequency of that frequency (N_r), that is, the number of sequences that have a certain subjective frequency r. The values N_r are then *smoothed*, that is re-estimated with a continuous downward-sloping line in log space. The smoothed values $S(N_r)$ are used to reestimate the frequencies according to (5.1):

$$r^* = (r+1)\frac{S(N_{r+1})}{S(N_r)} \tag{5.1}$$

The probabilities for frequency classes are then computed based on these reestimated frequencies:

$$p_r = \frac{r^*}{N} \tag{5.2}$$

where N is the total of the unnormalized estimates[3] .

Finally, the probability for unseen events is computed based on the (estimated) [4] probability of types of frequency one, with the following equation:

$$P_0 = \frac{S(N_1)}{N} \tag{5.3}$$

This probability P_0 corresponds to what I have called “propensity to generalize”.

As can be deduced from the equations, SGT is designed to ensure that the probability for unseen types is similar to the probability of types with frequency one. The propensity to generalize is therefore greater for distributions where most of the prob-

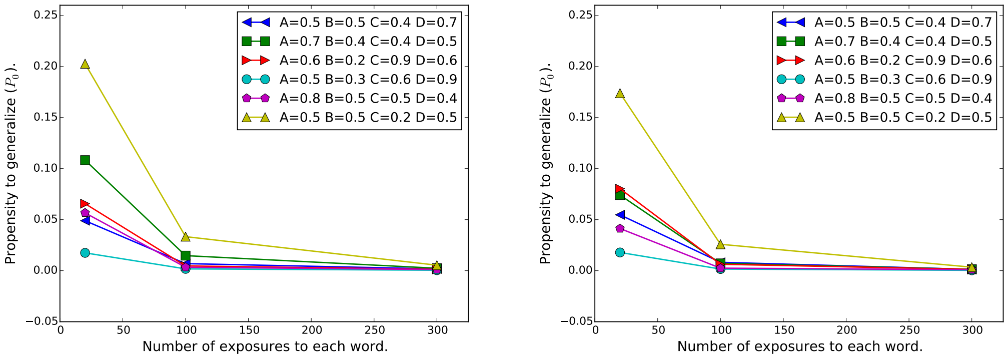
[3] It should be noted that the reestimated probabilities need to be renormalized to sum up to 1, by multiplying with the estimated total probability of seen types $1 - P_0$ and dividing by the sum of unnormalized probabilites.

[4] SGT incorporates a rule for switching between N_r and $S(N_r)$ such that smoothed values $S(N_r)$ are only used when they yield significantly different results from N_r (when the difference is greater than 1.96 times the standard deviation).

ability mass is for smaller frequencies. This obeys a rational principle: when types have been observed with high frequency, it is likely that all the types in the population have already been attested; on the contrary, when there are many low-frequency types, it may be expected that there are also types not yet attested.

5.5.2 Prediction of observed decrease in the propensity to generalize

Next, I apply the Simple Good-Turing method[5] to subjective frequencies computed by the R&R model. As shown in figure 5.4, the propensity to generalize (P_0) decreases as the exposure time increases, regardless of the parameter setting used in R&R. This result is consistent with the rationale in the Simple Good-Turing method: as exposure time increases, frequencies are shifted to greater values, causing a decrease in the smaller frequencies and therefore reducing the expectation for unattested sequences.



Exposure without pauses. Exposure with pauses.

Figure 5.4: Propensity to generalize, for several parameter settings (average of 100 runs). The model shows a clear decrease for all parameter settings, consistent with the empirical data (compare with figure 5.1).

The results of these simulations point to a straightforward explanation of the experimental finding of a reduced preference for the generalized sequences: longer exposures repeat the same set of words (and partwords), and consequently, participants may conclude that there are no other sequences in that language – otherwise they would have probably appeared in such a long language sample.

It can be noted in the graphs that the propensity to generalize is slightly smaller for the micropause condition. The reason for that is that R&R identifies words faster when micropauses are present, and therefore, the subjective frequencies tend to be greater.

[5] I use the free software implementation of Simple Good-Turing in https://github.com/maxbane/simplegoodturing.

This might appear unexpected, but it is in fact not contradicting the empirical results: as shown in figure 5.3, the difference between words and partwords is much bigger in the condition with micropauses, so this effect is likely to override the small probability difference (as would be confirmed by a model of step (iii)). It should be noted that, as reported in Frost and Monaghan [2016], micropauses are not essential for all type of generalizations (as is evidenced with the fact that rule*-words are generalized in the no-pause condition). Like those authors, I think the role of the micropauses is to enhance the salience of initial and final syllables (A and C) to compensate for the odd construction of the test items (rule-words and class-words), which include a middle syllable that occupied a different position in the familiarization stream.

5.6 Discussion

The experiments I have focused on are all based on the same simple language, but the results form a complex mosaic: generalization is observed in different degrees depending on the amount of exposure, the presence of micropauses and the type of generalization (rule-words, class-words or rule*-words). I have approached the analysis of these results with the use of several tools: first, with the three-step approach, a conceptualization of generalization that identifies its main components; second, with the use of R&R, a probabilistic model that already predicts some aspects of the results —and, importantly, generates a skewed distribution of subjective frequencies that is crucial for step (ii); and third, with the Simple Good-Turing method for quantifying the propensity to generalize. I now discuss how I interpret the outcome of my study.

Framing generalization with the three-step approach allowed us to identify a step that is usually neglected in discussions in ALL, namely, the computation of the propensity to generalize. I state that generalization is not only a process of discovering structure: the frequencies in the familiarization generate an expectation about the probability of next observing any unattested item, and the responses for generalized sequences must be affected by it. Moreover, this step is based on statistical information, proving that — contrary to the MoM hypothesis — a statistical mechanism can account for the negative correlation with exposure time.

It should be noted that this conclusion concerns the qualitative nature of the learning mechanism that is responsible for the experimental findings. It has been postulated that such findings evidence the presence of *multiple* mechanisms [Endress and Bonatti, 2016]. In my view, the notion of 'mechanism' is only meaningful as a high-level construct that may help researchers in narrowing down the scope of the computations that are being studied, among all the computations that take place in the brain at a given time. After all, there is no natural obvious way to isolate the computations that would constitute a single 'mechanism', from an implementational point of view. Therefore, the three-step approach should be taken as sketching the aspects that any model of generalization should account for, and the work reported here show that the experimental results are expected given the statistical properties of the input.

One issue to discuss is the influence of the use of the R&R model in computing the propensity to generalize. The Simple Good-Turing method is designed to exploit the fact that words in natural language follow a Zipfian distribution —that is, languages consist of a few highly frequent words and a long tail of unfrequent words. This is a key property of natural language that is normally violated in ALL experiments, since most of the artificial languages used are based on a uniform distribution of words (but see Kurumada et al. 2013). But it would be implausible to assume that subjects extract the exact distribution for an unknown artificial language to which they have been only briefly exposed. R&R models the transition from absolute to subjective frequencies, resulting in a distribution of subjective frequencies that shows a great degree of skew, and much more so than alternative models of segmentation in ALL. Thanks to this fact, the frequency distribution over which the SGT method operates (the subjective distribution) is more similar to that of natural language, and the pattern of results found for the propensity to generalize crucially depends on this type of distribution.

Finally, I have thus accomplished my goal qualitatively. The model captures the downward tendency of the propensity to generalize, but a model for step (iii), a long-standing question in linguistics and cognitive science, is required to also achieve a quantitative fit. The next chapters review the existing models for such step, and present a neural network model that reveal key properties of the generalization mechanism.

Part III

Generalization

Chapter 6

Computational Models of Rule Learning

This part of the dissertation addresses the third step of the conceptualization I proposed; namely, how individuals generalize to novel items. First of all, in this chapter[1] I present a critical review of models of generalization, and reflect on what is missing and which steps should be taken in future research. I then present a new model in chapter 7 that addresses some of the points raised in this review.

6.1 Introduction

One of the key abilities of human cognition is the capacity for discovering regularities in the environment and generalize them to novel cases. For instance, consider how, after seeing a giraffe for the first time, we are able to conclude that a previously unencountered giraffe is a member of the same animal species. Another well known example, introduced by [Fodor and Pylyshyn, 1988], is that an individual who understands the sentence *John loves Mary* can also understand *Mary loves John*[2]. To achieve that, an individual must abstract from the concrete properties in the input, and decide the extension of novel items to which the abstracted regularity or *rule* applies to.

In experimental linguistics, and concretely in the tradition of Artificial Language Learning (ALL), the study of how structural relations are generalized from language-like input is known as *rule learning*[3]. In the last two decades, a great body of experimental work has emerged, focusing on how humans discover relations that range from

[1] The content in this chapter is based on the following manuscript:
Alhama and Zuidema [2017c]. Computational Models of Rule Learning. [*To be submitted.*]

[2] Fodor and Pylyshyn refer to this property of human thought as *systematicity*. I see it as an instance of a generalized relation between two items (the *lover* and the *loved*).

[3] The term *rule learning* can be easily misinterpreted, since it has also been coined to postulate a specific theory of generalization (which I present later), according to which the knowledge extracted in such experiments must be explicitly represented as an algebraic symbolic rule. Unless otherwise specified, my use of *rule learning* is interchangeable with *generalization*; and *rule* is by default used to convey the same meaning as *pattern* or *regularity* unless it appears in the context of the specific theory that postulates algebraic rules.

identity rules [Marcus et al., 1999, Gerken, 2006, Endress et al., 2007] to nonadjacent dependencies [Peña et al., 2002, Gómez, 2002, Gómez and Maye, 2005, Endress and Bonatti, 2007, Frost and Monaghan, 2016] and even finite state grammars [Gomez and Gerken, 1999].

These experiments have inspired a number of computational models that aim to provide an explanation of the cognitive mechanisms underlying the empirical results. In this chapter, I focus on the experimental study presented in Marcus et al. [1999], which inspired a great number of modelling approaches and a very active debate around theoretical issues such as the nature of the representations involved.

This chapter is structured as follows. First I describe the empirical findings from the experiment by Marcus and colleagues (§ 6.2). Then I describe the existing models (§ 6.3, § 6.4), to then identify what are the relevant questions that the models address (§ 6.5). Finally, I delineate an agenda of concrete desiderata for future modelling efforts (§ 6.6).

6.2 The Empirical Data

Marcus et al. [1999] presented an ALL study that aimed to investigate the acquisition of grammar-like rules by 7 month old infants. The authors run a total of three experiments, in which the 7 m.o. participants are familiarized with a speech stream consistent with a certain 'grammar' or pattern. For instance, one of the familiarization streams could involve syllable triplets like "linali talata nilani gagiga ..." (where the spaces denote silence gaps); in this case, we say that the stimuli were generated with an ABA grammar. In order to see if infants learn to extract the grammatical rule, they are subsequently tested with stimuli that involve triplets consistent with the grammar they were familiarized with, and triplets generated with another control grammar. The amount of time that infants direct their attention to the stimuli being played (the 'listening times') are recorded, and then a statistical test is applied to see if they are significantly different between stimuli of the different grammars, showing in this case that infants learnt to discriminate stimuli of each grammar. Crucially, the test stimuli contain syllables that did never appear in the familiarization stream; in this way, infants cannot solve the task by only memorizing the syllables they were familiarized with.

In the first experiment, half of the participants are assigned to one of two conditions, differentiated by the grammar used to generate the familiarization stimuli: either ABA or ABB. After the familiarization phase, infants are tested to see if they extracted the underlying grammatical patterns. In accordance to the procedure described above, for all the participants the test stimuli contain triplets from both the ABA and the ABB condition. The complete stimulus set is listed in table 6.1.

A statistical analysis of the results of this first experiment shows that looking times were significantly longer for inconsistent triplets, reflecting what the authors interpret as a novelty preference. It seems therefore that infants are able to discriminate between grammatical and ungrammatical items. Since the syllables in test and familiarization do not overlap, the results seem to be an indication that the infants have abstracted the grammatical rule.

	Familiarization				Test
ABA	ga ti ga	li ti li	ni ti ni	ta la ta	wo fe wo
	ga na ga	li na li	ni na ni	ta ti ta	de ko de
	ga gi ga	li gi li	ni gi ni	ta na ta	
	ga la ga	li la li	ni la ni	ta gi ta	
ABB	ga ti ti	li ti ti	ni ti ti	ta la la	wo fe fe
	ga na na	li na na	ni na na	ta ti ti	de ko ko
	ga gi gi	li gi gi	ni gi gi	ta na na	
	ga la la	li la la	ni la la	ta gi gi	
3x triplet (random order)					

Table 6.1: Stimuli used in experiment 1 in Marcus et al. [1999].

	Familiarization				Test
ABA	le di le	wi di wi	ji di ji	de di de	ba po ba
	le je le	wi je wi	ji je ji	de je de	ko ga ko
	le li le	wi li wi	ji li ji	de li de	
	le we le	wi we wi	ji we ji	de we de	
ABB	le di di	wi di di	ji di di	de di di	ba po po
	le je je	wi je je	ji je je	de je je	ko ga ga
	le li li	wi li li	ji li li	de li li	
	le we we	wi we we	ji we we	de we we	
3x triplet (random order)					

Table 6.2: Stimuli used in experiment 2 in Marcus et al. [1999].

However, the authors identified a possible confound: the consonants in the stimuli appear always in voiced – unvoiced – voiced combinations, so the results could potentially be explained also as caused by the detection of such a pattern. In order to rule out this possibility, a more carefully controlled replication of the previous experiment is presented (the list of stimulus items is shown in table 6.2). The responses in this experiment still prove to be statistically significant in discriminating between grammars, and therefore, the conclusions are consistent with those of experiment 1.

Finally, there still exists the possibility that the infants focus on the presence or absence of an immediate repetition, rather than on the abstract identity rule. Therefore, the authors carry out a third experiment in which the stimuli grammars are ABB and AAB. Again, the participants show significantly different responses between grammars, so this alternative explanation is ruled out.

To summarize, the three experiments show significantly different responses for each of the tested grammars, so it seems that infants find some regularity that allows them to discriminate between grammars. This discrimination happens even when phonetic features are more carefully controlled, and it goes beyond the simple presence or absence of an immediate repetition, so the authors conclude that 7 m.o. infants must be abstracting the identity rule in the stimuli.

6.3 The Neural Network Models

In light of the reported results, Marcus and colleagues reflect on the cognitive mechanism that may be responsible for the results. The first option they consider is whether the same mechanism attested for word segmentation (*statistical learning*, as explained in previous chapters [Saffran et al., 1996a,b, Aslin et al., 1998]) could be at play during this experiment. However, Marcus and colleagues conclude that the results are incompatible with such an explanation, since the statistics for novel items amount to zero. Instead, the authors propose that a cognitive mechanism of a different nature must be at play; concretely, a *rule-based* mechanism that extracts *algebra-like rules* over *variables* – that is, a mechanism that explicitly incorporates operations over *symbols*.

In order to provide additional support for this idea, the authors report failed simulations with a neural network architecture, a class of model that implements *statistical learning*. The authors argue that neural network models like the one they simulated (a Simple Recurrent Network or SRN, Elman [1990]) do not stand a chance to account for the results precisely because such models do not explicitly encode variables and relations over variables. The scope of this claim includes the most standard neural network architectures, which do not incorporate symbolic operations; but it did not target hybrid neural network approaches that are extended to represent variables.

This study received much attention; after all, what Marcus and colleagues pointed out was an apparent limitation of neural networks in reproducing a very basic capacity that does not seem to pose a problem for infants, already at a very young age. The publication triggered a heated debate, in which connectionist modellers presented a number of alternative neural network models to account for the results; the next sections are devoted to reviewing these models.

6.3.1 The Simple Recurrent Network Models

The Simple Recurrent Network (SRN) was proposed in Elman [1990] as a variant of the classic feed-forward network, specialized in learning regularities over *sequences*. The main contribution of this network is to process input data that unfolds through time. The architecture of an SRN (which can be seen in figure 6.1) incorporates a ‘context’ layer. At every time step, the activation values of the hidden nodes are copied into the context layer. The context layer is at the same time connected as an input to the hidden layer, so the hidden layer reads activations from both the input layer and the context layer. Therefore, the internal representations depend on the previous state of the network, incorporating in this way a form of memory.

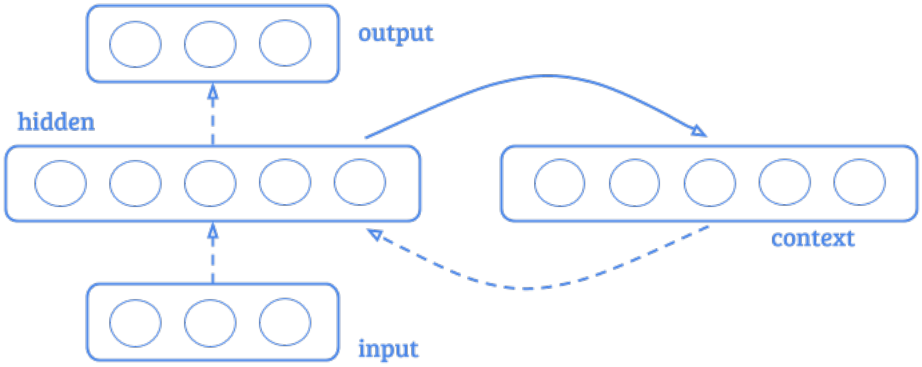


Figure 6.1: The Simple Recurrent Network [Elman, 1990]. Dotted lines represent trainable connections; continuous line represents a one-to-one copy of the activation values.

As mentioned before, Marcus and colleagues report failed simulations with this model; however, other researchers proposed to model the experiment with an SRN with certain modifications. I now review these models.

Using analog encoding

Negishi [1999] reflects that the lack of generalization in the original simulations is due to the fact that both of the encoding schemes employed are based on *binary* activations. Instead, the author proposes to represent features of the input as real numbers.

The original simulations were replicated, but with the stimuli characterized by two continuous (analog) values: vowel height and place of articulation. This resulted in a model that produced larger prediction errors for the grammar it was not trained with, a fact that may be interpreted as reproducing the increased attention over inconsistent test stimuli observed in the experiment.

Marcus [1999c] argues that the use of analog encoding can be seen as endowing the network with registers: if an input node represents all possible values, then it suffices to connect it to the output node with a weight of 1, and thus, the node would act like a variable that instantiates a particular value at a given time. However, this reasoning is not so straightforward when it applies to SRN models: the non-linearities in the hidden layer, and the connection with the recurrent layer do not permit the direct copy proposed by Marcus. As argued in Sirois et al. [2000], variable bindings are only effective if they can be accessed for further computation.

Optimizing for a different goal: Segmentation, Categorization and Transfer Learning

Christiansen and Curtin [1999] (and later also Christiansen et al. [2000]) suggest that the statistical knowledge that infants acquire when attempting to segment the speech in their environment could be the basis for their success in the experiment reported by Marcus and colleagues. To prove this point, they use an existing SRN model that learns to segment speech using different types of probabilistic information [Christiansen et al., 1998].

Their model is presented with a sequence of phonemes (instead of syllables), encoded with phonological features, primary and secondary stress, as well as whether the phoneme is the last one in a triplet (and therefore it is followed by a 1s silence gap; see table 6.3 for more details). The model is trained to predict an arbitrary representation of the next phoneme in the sequence, but also whether the phoneme is a syllable boundary, that is, whether it is followed by a 250s silence gap (which is the length for pauses *within* triplets). In this way, the model is expected to learn to segment syllables after having been given the information of triplet boundaries.

In order to evaluate the performance of the model, the authors introduce two novel methods. First, they report that the network performs better at segmenting syllables belonging to triplets that are not consistent with the training grammar. The authors interpret this as accounting for the behavior of infants in the experiment, who pay more attention to inconsistent test items. Second, an analysis of the internal representations built by the model is performed. The authors find that the representations for consistent and inconsistent triplets are distinguishable, as revealed by a two-group discriminant analysis.

Marcus [1999c] argues that an analysis of the internal representations is not a suitable evaluation, since representations must have a causal effect on the output in order to be meaningful. Although the segmentation task could potentially account for that, Marcus observes that it is somewhat unnatural that the model is trained with stimuli that represents the pauses between triplets while being tested on the type of pauses that have intentionally not been coded. Additionally, Marcus observes that the statistical significance of the analysis of the internal representations may not be meaningful, since the test consists of a very small number of items (4) compared to the number of hidden units (80) that provide the internal representation.

* * *

The next model I review (Altmann and Dienes [1999], based on an earlier model by Dienes et al. [1999]) conceives of generalization as an instance of transfer learning between different domains. In the context of the Marcus et al. experiment, the authors identify the domains as defined by familiarization stimuli and test stimuli. In order to account for the distinct domains, the authors extend the SRN architecture; concretely, the input and the output layers are augmented with extra nodes, such that two separate groups of nodes in each layer account for each domain. Additionally, the SRN is extended with an extra layer (the "encoding" layer), situated between the input and the hidden layer. The architecture of this network can be seen in figure 6.2.

The network is first trained as a normal SRN, using only the input and output nodes of the first domain (D1). In the test phase, the stimulus is presented to the

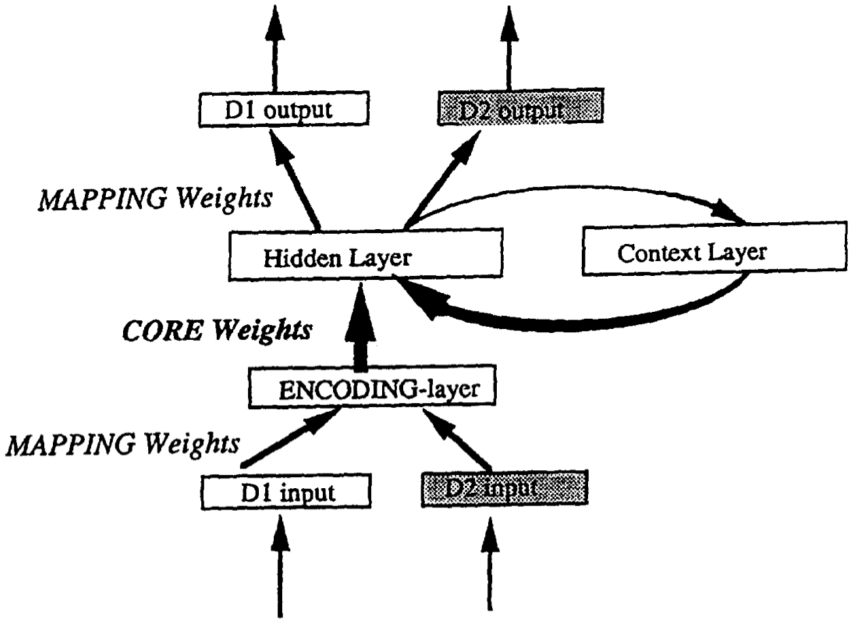


Figure 6.2: SRN model in Altmann and Dienes [1999] (image from Dienes et al. [1999]; copyrights remain with the original holders).

group of input nodes corresponding to the second domain (D2). Crucially, the test items are presented several times, and – contrary to the previously reviewed models – the network continues updating the weights, with the exception of those connecting the encoding layer and the hidden layer, which remain "frozen". By keeping those weights intact, the model preserves some of the knowledge learnt during training and attempts to transfer it to the test stimuli.

The authors measure the success of the network by computing the Euclidian distance between the predicted and target vectors, for the vectors resulting from the last iteration in the test. The results show higher correlations for prediction in the consistent grammar. However, Marcus [1999e] observes that this implementation is consistent with the experimental results only when evaluating the results by computing the distance to the target. If, instead, one evaluates the most active unit in the predicted vector, then the model oscillates between the two grammars.

Additionally, it must be noted that in this model the domains are predefined, and there is no a priori reason for the test items to be part of a different domain. Accordingly, the model requires a mechanism that selects and freezes a subset of the weights, but it is not clear when and under what circumstances this mechanism would operate, and which particular subset of weights it should freeze. As Marcus puts it, the model is task-specific, and it remains unclear how it can be related to other cognitive mechanisms.

* * *

I now review a model that deviates from the original simulations in two ways: by changing the task into *categorization*, and by accounting for prior experience. Seidenberg and Elman [1999a] observe that the SRN presented by Marcus et al. had no previous knowledge, while infants in their experiment had been exposed to natural language in their environment. The authors argue that, by this prior exposure, infants might have learned to represent phonological similarity between syllables.

In order to account for prior knowledge, the authors extensively *pre-train* an SRN with 120 different syllables. In this pre-training phase, one single node is optimized to output whether the current syllable is the same as the previous syllable in the sequence. In this way, the SRN is trained to learn *identity* between syllables.

The weights learned during pre-training are used to initialize the SRN for the actual experiment, which is also defined as a categorization task, this time involving a different output node. Crucially, the network is not trained only with items belonging to one type of grammar (as the infants in the original experiment), but also with triplets generated from *both* ABA and ABB.

When tested with the novel triplets, the network shows responses close to zero for the ABA triplets and closer to 1 for ABB (concretely, 0.004 and 0.008 for bapoba and kogako, and 0.853 and 0.622 for bapopo and kogaga). Thus, it seems that the SRN learnt to correctly discriminate between the grammars.

However, although the incorporation of pre-training is cognitively motivated, other aspects of this work require further justification, as discussed also in response letters Marcus [1999c], Seidenberg and Elman [1999b], Marcus [1999d]. The simulations greatly deviate from the original experiment in providing the model with negative evidence, and additionally, the incorporation of a feedback signal both during pre-training and training does not have its counterpart in the original experiment, since subjects in the experiment did not receive any form of feedback.

Marcus further observes that the output node is trained to follow the symbolic rule ‘if X==Y then 1 else 0’, suggesting that this evidences the need for symbolic operations. Although the model is clearly trained under that rule, as Seidenberg and Elman argue, the feedback is an external signal, which does not modify the space of hypothesis of the model. In other words, the fact that the supervision signal can be expressed with a symbolic rule does not entail that the network implements symbolic operations. It is nevertheless not clear where the signal for learning identity on the first place would come from, and whether it is plausible that a region of the brain is dedicated to finding identity relations in the input.

Accounting for previous experience

I now review a model that also incorporates prior experience, but in this case it remains as the original prediction task.

Altmann [2002] simulates the experiment in the same architecture presented in Dienes et al. [1999] and Altmann and Dienes [1999] (reviewed in section 6.3.1). Like in the previous proposals, the model is allowed to learn during the test phase; however, in this study, no connections are frozen — that is, all the connection weights of the model can be updated.

The pretraining consists of a prediction task over a set of 252 sentences in natural language, which were generated from the grammar and vocabulary presented in Elman [1990] (with an additional control experiment that avoids sentences involving ABA or ABB structures). After the pretraining, the network is trained on the Marcus et al. stimuli, and then tested with the novel items.

The output of the model is evaluated by computing the product moment correlation between the predicted vector with the correct one. Statistical tests show that the response of the model significantly varies between consistent and inconsistent items. The authors conclude that their model reproduces the empirical data; however, the critiques that Marcus raised for previous work [Marcus, 1999e], based on the fact that the model iterates over the test items several times, apply also to these simulations.

Additionally, in my opinion a relevant aspect in this model is the type of representation employed for pauses. The authors use two different vectors to encode the pauses: one for pauses that precede the onset of a triplet, and another to mark the ending. In this way, the learning process can exploit this information to detect different associations for onset and final syllables. This constitutes an indirect form of positional information that is not available in the actual familiarization stream (since, perceptually, there is a single uniform silence gap between triplets).

6.3.2 Neural Networks with non-sequential input

I now review two neural network models that are based on a different architecture, known as *autoencoder* or *autoassociator* [Mareschal and French, 1997]. These models do not incorporate recurrent connections to learn sequential relations over the input; instead, they are optimized to find a suitable *encoding* of the input. The objective function is set to minimize the error in reproducing the input pattern in the output, so the network needs to built intermediate representations that favour the reconstruction of the input.

I first review the study presented in Shultz [1999] (later replicated in Shultz [2001] with a different encoding scheme). This model is an autoencoder that is trained with *cascade-correlation* [Fahlman and Lebiere, 1990]. The main property of cascade-correlation is that it gradually adds new nodes to the hidden layer, as determined by the computed error. Since the network is trained to reproduce the input in the output layer, the error signal is based on the difference between the actual output and the input.

In the original simulations, each syllable is encoded as a real number, which is represented in a single node. The information is presented to the network as triplets; thus, the input and output layers consist of 3 nodes, each one corresponding to one syllable.

For the evaluation of the results, the author submits the error produced in the output layer to a repeated ANOVA. The results show a significant effect on grammar condition, with more error for inconsistent test items. Shultz concludes that this reproduces the original experiments, since more error requires further cognitive processing –as would be reflected in the increased looking times for inconsistent items.

Marcus [2001] argues that, due to using just one node to represent each syllable, the model can easily learn to copy the relevant syllables in the output. It is true that given the topology of the network, which is is built *on the go*, it is easy to imagine that with the incorporation of a few nodes, the input gets roughly copied in the output (although distorted with the non-linear function applied to the hidden nodes), but Vilcu and Hadley [2005] showed with further analysis that the network does not perform such mapping.

Vilcu and Hadly further argued that these results are only replicable for this particular stimuli; for ABA or ABB sequences that involve different phonemes, the model is unable to distinguish between the grammars. Additionally, the authors show that the model does not generalize to stimuli encoded outside the range of the real numbers employed in the encoding.

It must be noted that this network operates over full triplets, creating in this way a somewhat artificial treatment of a continuous input. The next model I review incorporates a slightly more realistic treatment of time.

* * *

Sirois et al. [2000] implement a fully connected neural network; that is, the nodes in the network are all connected to each other. The network is trained to reproduce external input in its nodes, strengthening the connections in the nodes that have correlated activations.

This model contrasts with the previous approaches in its formalization of time. Each syllable in a triplet is presented to a different group of nodes, one at time. After the three syllables in a triplet have been presented, the activations are reset. Additionally, the model implements activation decay. For instance, when the third syllable of a triplet is presented, the first and second syllables still remain active, but their activations are decreased (with the first syllable having weaker activation than the second).

As for the evaluation, Sirois analyses the number of presentations required for the test items to be assimilated (i.e. to be accurately reproduced by the network). The results of this analysis exhibit a significant difference between consistent and inconsistent items; therefore, the authors interpret the results as showing that this model could be a possible explanation of the original experiment.

It must be noted though that one of the drawbacks of this approach is that its design is intimately tied to the actual experiment: the architecture would need to be adapted in the case of input sequences with a different number of syllables, and the full activation reset is linked to the appearance of the highly salient 1s silence gaps in the input, but it is not clear how this would generalize to stimuli involving less perceptible silence gaps (if any).

6.3.3 Neural Networks with a repetition detector

We now review three models that include some form of dedicated mechanism to detect repetitions of syllables in the input.

Shastri and Chang [1999] present a model of the Marcus et al. experiment that implements a form of dynamic binding [Hummel, 2011]. The model, originally presented in Shastri et al. [1993], is implemented as a neural network with two groups of dedicated nodes for the input: one group that represents the phonetic features for an input syllable, and another group with three nodes corresponding to each of the three positions in a triplet. The idea behind dynamic binding is that the nodes of the two groups activate in synchrony, and therefore the *coincident* activity can be exploited by the network in order to learn the abstract pattern in the input.

This neural network model involves recurrent connections, but it is not implemented as a standard SRN. Instead, the model (illustrated in figure 6.3) clamps some input activations and propagates the activity through the network. After some delay, a target (the "correct" activation of the positional information) is clamped in the network. The difference between the actual activations and the target is used to update the weights through gradient descent.

Crucially, during the presentation of each syllable, all the positional nodes in which the syllable appears are active in the target; for instance, for the stimulus *ledile*, the first and the third positional nodes are *both* active on *each* presentation of *le*, and the second positional node is active during the presentation of *di*. It must be noted that, with the introduction of this form of feedback, the model is provided with an actual mechanism for detecting repetitions.

When it comes to the Marcus et al. experiment, the performance of the model is evaluated by computing the mean squared error between the model activations of the positional nodes and the target. The error is considerably smaller for test items consistent with the training grammar, and thus the model appears to reproduce the empirical findings. However, this type of evaluation is based on the actual relations learnt by the model (since it employs a target based on a rule), rather than on the produced output (the "behaviour") of the model. Therefore, it cannot be compared to the results in the experiment with infants without assuming that these are the rules extracted in the experiment.

Shastri and Chang argue that this approach offers a plausible mechanism to implement *rules* via biologically inspired temporal synchrony. Thus, this model is not presented as a counterargument to the claim by Marcus et al.; actually, Marcus [2001] reflects that this model implements *temporal* rather than *spatial* variables. However, as argued also in Shultz [2001], the design of the model is very tied to the actual experiment; additionally, the feedback is clearly unrealistic, in providing the model with the expected outcome rather than with the available information in the input. Therefore, it does not really offer a reconciliation between symbolic and neural network models.

* * *

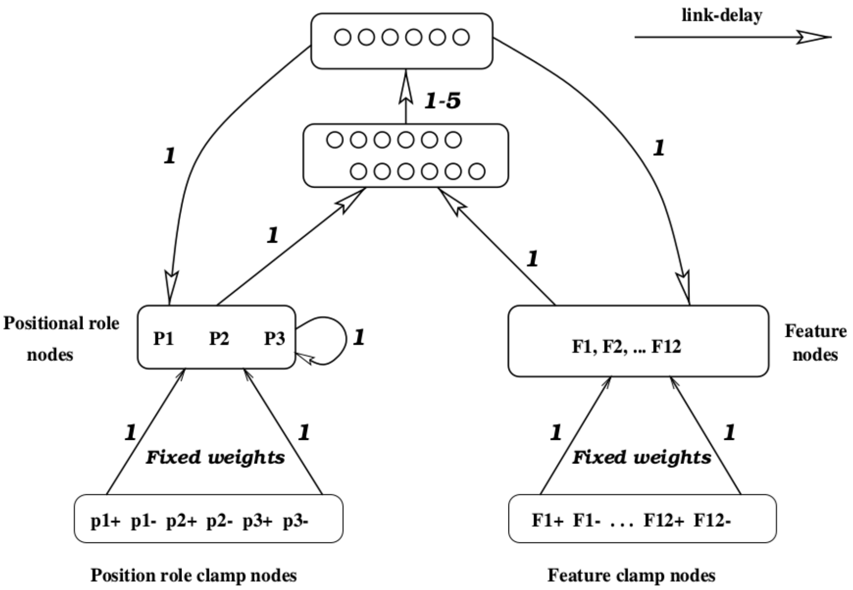


Figure 6.3: Model presented in Shastri and Chang [1999] (image courtesy of the authors; copyrights remain with the original holders). The numbers indicate link delays in the connections.

Gasser and Colunga [2000] also present a model that implements another form of dynamic binding. The authors frame the problem as extraction of correlations from the input; in their view, the reason why the correlations that infants learn during the experiment allow them to generalize to novel items is that those correlations are *relational* instead of content-specific.

This model – named PLAYPEN – is implemented as a generalized Hopfield network, that is, a fully connected neural network model in which weights are adjusted with the Contrastive Hebbian Learning algorithm [Hopfield, 1984]. The network, illustrated in figure 6.4, is provided with dedicated units that detect sameness and difference. Therefore, its task is to reinforce the correlations according to the relations of sameness and difference found in the input.

As in the previous model, the authors assume dedicated nodes for each syllable position in a triplet. Additionally, the model is augmented such that each unit has an “angle”. While the particular value of the angle is irrelevant, it provides an additional dimension to the network, such that units with similar angle can be treated similarly. In this way, Gasser and Colunga implement a form of simplified dynamic binding.

The results of the experiment simulations with this model show that the relational units are more active for test items consistent with the training grammar. Therefore, the network has strengthened the connections of the relations present in the familiarization stimuli.

In spite of the claims of biological plausibility of the model, its actual implemen-

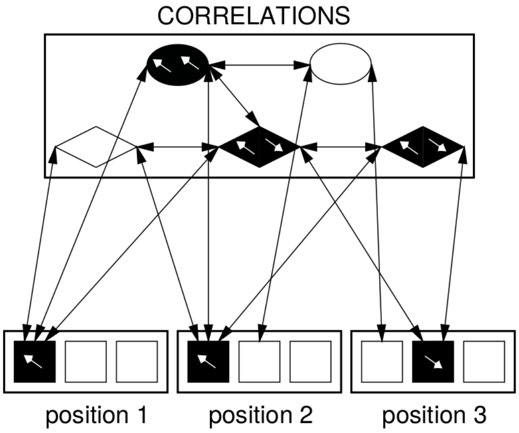


Figure 6.4: The PLAYPEN model [Gasser and Colunga, 2000] (image courtesy of the authors; copyrights remain with the original holders). Diamonds indicate difference relations; ovals indicate sameness.

tation remains extremely tied to the actual task, since the model comes with pre-wired relations over explicit bindings. Still, the authors argue that their model does not qualify as a symbolic model, since variables in symbolic models are content-independent, while PLAYPEN is sensitive to feature similarity between the presented items. However, while I agree that the model does not implement symbolic variables, it does incorporate rules, and thus it embodies the assumption that infants are equipped with a repetition detector.

* * *

The last model in this section is a variant of a recurrent neural network model that also incorporates a repetition detector. This approach, presented in Dominey and Ramus [2000], is based on an architecture called Temporal Recurrent Network [Dominey, 1995], a recurrent architecture in which only the weights connecting the hidden and the output layer are trained (while the rest are randomly initialized and remain unchanged). Interestingly, the nodes in this network are Leaky Integrate-and-Fire units. This endows the network with a more realistic representation of time: unlike in previous approaches, in which activations consist of a single value that is produced after a discrete timestep, the nodes of this network produces activations in the form of continuous *spikes*.

Nevertheless, the authors do not find with this model a pattern of results that is consistent with behaviour of the infants in the experiment. Therefore, they explore an augmented version of the model, which they call Abstract Recurrent Network (ARN). The ARN model features an additional short-term memory that stores the last 5 syllables of the input (in order to account for the 7±2 *magical number* for short-term memory, [Miller, 1956]), and a "recognition" component that detects whether any of the items in the short-term memory is the same as the actual item. This information is then provided to the internal state (the hidden layer), so that the model can exploit this information while updating the weights between the hidden layer and the output.

Given that the model is updated in continuous time, the responses can be easily compared to Reaction Times. Concretely, the authors measure the response latencies of the activation of the correct output nodes. These latencies should be smaller for learnt items, since the strength of the activations in the network would influence the activity in the output. With this form of evaluation, the authors find significant statistical evidence for shorter Reaction Times for test items consistent with the familiarization grammar.

This model therefore appears to be a promising approach, in incorporating successful learning with a more realistic treatment of time. However, as mentioned in the previous approaches, this network is endowed with a component that actively looks for repetitions in short-term memory (and even contains a node that fires when no repetition is found). This results in the very strong postulation that infants must be equipped with such dedicated mechanism.

Importantly, this model adds some form of variables, in line with the claims by Marcus and colleagues. This is due to how the model accesses the augmented short-term memory, which is based on isolated memory positions dedicated to store different items. It must be emphasized that this behaviour comes from how the memory is *accessed*, not by the mere fact of adding short-term (spatial) memory; in other words, a network with this kind of memory may *learn* how to access it efficiently, and might eventually discover how it can exploit that nodes in the memory are dedicated to different items. This is not the case in the model by Dominey and Ramus, in which the recognition component is handcrafted to have positional access to elements in the short-term memory.

6.3.4 Evaluating the Neural Network Models

Although the models reviewed address the same experimental data, they greatly differ in the criteria applied to judge their success in reproducing the empirical phenomenon. In the original simulations by Marcus and colleagues, it is not clear how the output of the model was evaluated, since the authors only report that "the network is unable to distinguish the inconsistent and consistent sentences". In the follow-up papers, modellers use some error measure (e.g. Mean Squared Error) to compute how much the predicted vectors deviate from the target [Negishi, 1999, Altmann and Dienes, 1999, Altmann, 2002, Shultz, 1999, Shastri and Chang, 1999]; generally this is accompanied with a statistical analysis that shows whether the computed error allows for distinction between grammars.

While that is the most common approach, the rest of the reviewed studies apply completely different criteria. To begin with, Seidenberg and Elman [1999a] report the average activity of the output node in their model (which is trained for grammar classification). Another example is Sirois et al. [2000], who compare the number of stimuli presentations required to succeed in generalization in each grammar. We find yet another approach in Gasser and Colunga [2000], who analyze whether the model learns the expected identity rule (an approach that is also present in an extended analysis in Shastri and Chang [1999]). In contrast, Christiansen and Curtin [1999] and Christiansen et al. [2000] evaluate their model in its prediction of the internal pauses within a triplet (since the authors frame the problem as a segmentation task). Additionally, the authors also analyze the internal representations built by the model, and find that they are distinguishable for each type of grammar. Finally, a contrasting approach is found in Dominey and Ramus [2000], in which the latency of output node related to the correct prediction is measured and taken as an indication of Reaction Time.

Given this stunning variation on evaluating the models, it is not possible to quantitatively compare the models in either their goodness of fit to the empirical data or in to what extent the models are capable of generalization.

Interestingly, there is one aspect of the evaluation criteria on which most of the proposals agree: the models are evaluated in showing some form of *discrimination* between grammars. The corollary is that, even though a successful model can reliably detect some difference between the stimuli generated by each grammars, it need not be capable of accurately *generalizing* (i.e. on systematically predicting or accepting novel items that are consistent with the generalized rule). Exceptions to this are the models by Shastri and Chang [1999] and Gasser and Colunga [2000], which are evaluated on learning abstract identity (or difference) relations –albeit they are not evaluated on producing actual generalizations but rather on settling for the hypothesized rules.

6.4 Symbolic Models of Rule Learning

All the models reviewed so far are neural network approaches, albeit some of them incorporate some form of symbolic structure. The original proposal by Marcus and colleagues can also be implemented in a purely symbolic model, but the number of such models applied to the Marcus et al. experiment is really small (to the best of my knowledge, the only existing models are Kuehne et al. [2000] and Frank and Tenenbaum [2011]). A possible reason for this asymmetry with neural approaches is that the grammars to be learnt are so simple when rules and variables are assumed that there is not much room for unexpected findings. For this reason, here I focus on the model by Frank and Tenenbaum [2011], since its the only model that goes beyond simple rule extraction to also address an important question that has been neglected in other approaches; namely, what determines which rule is preferred when there are multiple consistent rules.

The model presented by F&T is a Bayesian model, inspired by the Bayesian similarity model in Tenenbaum and Griffiths [2001]; however, unlike the original proposal, it is not driven by analogy between exemplars; instead, the model infers the most likely symbolic rule that may have generated the observed data.

The hypothesis space in this model consists of an inventory of symbolic rules. Since the model is applied to a variety of datasets, this inventory is adapted to each experiment; in the case of the experiment by Marcus and colleagues, the hypothesis space incorporates an identity relation over syllables. Therefore, the model may learn any combination of identity relations over syllables in the input.

The model uses Bayesian inference to decide which is the most likely rule r (grammar) that may have generated the observed stream of triplets $T = t_1, ..., t_n$; in other words, the goal of the model is to find for which rule r the *posterior* probability $p(r|T)$ is maximal. As defined in equation 6.1, the posterior can be computed as a product of the *likelihood* of observing the data assuming this rule $p(T|r)$ (see equation 6.2) and the *prior* probability of the rule $p(r)$ (that is, the probability of a rule before observing the input stream). In order for the posterior to be a probability, this product needs to be normalized by the sum of the probabilities for all possible rules.

$$p(r|T) = \frac{p(T|r)p(r)}{\sum_{r' \in R} p(T|r')p(r')} \tag{6.1}$$

where

$$p(T|r) = \prod_{t_i \in T} p(t_i|r), \tag{6.2}$$

and where

$$p(t_i|r) = \frac{1}{|r|}. \tag{6.3}$$

The prior is defined as a uniform distribution; hence, each rule is *a priori* equally likely. As for the likelihood, the authors assume *strong sampling*, that is, the triplets are assumed to have been uniformly sampled from the set of triplets that a rule can generate. This entails that the probability of observing a triplet under a certain grammar is larger for "smaller" grammars (that is, grammars that generate a smaller number of triplets), as seen in equation 6.3. This creates a bias in the model in favour of more concise grammars; for instance, for a triplet such as "je-li-je", a rule like ABA is more likely than a more general but equally consistent rule ABC that involves no identity.

Given the simplicity of the experiment and the model, the only rule that can compete with ABA or ABB is ABC, that is, a grammar which generates triplets consisting of three arbitrary syllables, which may or may not be repeated. This model shows that more probability is attributed to the more specific grammars ABA and ABB (consistent with the *size principle* defined above).

The authors present two additional variants of this model: one that assumes a certain amount of noise in the generative process (regulated through an additional parameter), and another that additionally allows for the possibility that the data was generated from multiple rules.

Due to the addition of parameters, the model now requires a procedure of *fitting*. The posterior probabilities derived from the model are related to the human responses through the use of the negative log probability (surprisal). These two models also attributed more probability (less surprisal) to the rules involving identity, although some small probability mass was attributed to the general rule ABC.

Thus, the BLM model identifies which rules would be favoured under a rational model incorporating a certain bias for less general rules. Although the model appears extremely simple in comparison to the neural network approaches (mostly due to the fact that that the hypothesis space is relatively small and manually defined, *ad hoc* for the experiment), it brings an additional value that should not be underestimated. Concretely, it is the only model that clearly defines which biases are hypothesized to guide the preference for certain rules over others. Thanks to that, this approach incorporates a principle that postulates why the participants induced an identity rule instead of a more general rule in which all triplets are possible.

This study was fiercely criticized in a follow up publication [Endress, 2013]. Among other issues, the author questions the validity of the "size principle" as a cognitive bias. The author reports an independent experiment in which participants are exposed to instances of an ABB grammar, vocalized with human speech. The participants could discover at least two rules: the identity rule between the second and third syllable, or a more general rule glossed as "any sequence of human speech syllables". In the subsequent test phase, participants had to choose between an ABB sequence of monkey vocalizations, or AAB triplets carried by human speech. The results show a preference for the AAB human triplets, a fact that is interpreted by Endress as contrary to the "size principle". However, Frank [2013] argues that this data shows a modality preference rather than a rule preference. The issue is therefore not settled: it is not clear whether the subjects prefer certain generalizations based on the "size" of the set of items that a rule can generate or based on a bias favouring acoustic aspects of the input (i.e. the vocalizations).

6.5 Analysis of the Models

In the previous section, I have reviewed models that appear to offer distinct perspectives on generalization. I now identify what I think are the most relevant questions that the Marcus et al. experiment give raise to, and analyse whether these seemingly contrasting approaches differ in answering those questions.

6.5.1 Question 1: Which features or perceptual units participate in the process?

In the experiment, infants are exposed to a synthesized speech stream. The way this stream is perceived must impact what is learnt from it, and therefore, details of this perceptual process are relevant. For instance, do infants perceive the input as a sequence of phonemes, or is the syllable the most salient perceptual unit? Do they analyze lower-level properties, such as phonetic features, once a syllable has been recognized? Do other acoustic dimensions, such as loudness or pitch, play a role in what is learnt? How does the insertion of silence gaps affect the perception of the basic units?

These questions have not received much attention. However, thanks to using computer modelling, researchers are forced to make choices on how to represent the input and how to present it to the model over time. This is reflected in the encoding, which may, for instance, incorporate some detailed acoustic aspects of the stimuli or, instead, represent it with arbitrary symbol that does not encode any properties of the item. The latter is the approach taken by symbolic models such as the BLM, where each syllable is represented with an arbitrary symbol that does not incorporate any information of what it represents. But in the case neural network models, the input vector can be coded in either way, as explained next.

In a localist encoding scheme, vectors are initialized to a null value (typically, 0 or -1) except for one of its units, which will have a non-null activation value (typically, 1). The position of the active unit indicates which item is represented, but it does not convey any information of the properties of the item; therefore, localist representations are always arbitrary. On the contrary, in a distributed encoding scheme, each unit may participate in the representation of more than one item. Although these values can be chosen arbitrarily, each unit may be chosen to represent one particular property of the item, and thus, the distributed vector would encode certain specific properties of the stimuli. As shown in table 6.3, models vary in the choice of represented acoustic features.

The choice of the encoding scheme has an impact on generalization. As argued in Marcus [1998], for a neural network to succeed in generalizing to novel items, such novel items must fall *within* the training space; that is, the values of each unit must have been witnessed by the network during training (even if in different combinations). For instance, if the training data contains the vectors [1, 1, 1], [1, 0, 1] and [1, 0, 0], then a novel item like [1, 1, 0] lies within the training space, while [0, 1, 0] is outside the training space, since the first unit in the vector has never taken the value 0 during training. If a certain unit in the input always has the same value during training the network learns a solution based on such fixed value, and therefore the solution will not be valid for novel test items that contain the opposite feature value.

For this reason, the choice of the encoding scheme and the dimensions to encode is relevant, since a certain amount of overlap is needed for generalization in neural networks. To illustrate this, consider the case in which localist representations are used. The nodes that represent the test items will be zero during familiarization, and therefore the learning algorithm will update the connecting weights until they converge to zero, so they would never be active to predict novel items. On the contrary, with distributed vectors, some of the units representing the test stimuli may have been active during training.

This raises two relevant issues. First of all, distributed vectors stand a chance for generalization, depending on the overlap between vectors in training and test. This can be accomplished in two ways: either by using pseudo-random symbolic initializations that guarantee a certain amount of overlap, or by investigating which are the relevant properties of the input that need to be coded in the vector. To our knowledge, this issue has not been thoroughly explored. Thus, this will be one of our points in the desiderata

Models	Unit	Scheme	Features
Marcus et al. (1))	Syllable	Localist — Binary	
Marcus et al. (2)	Syllable	Distributed — Binary	6 phonetic features
Negishi	Syllable	Distributed — Analog	Place of Articulation and Continuous Vowel Height
Christiansen&Curtin	Phoneme	Distributed — Binary	11 phonological features, primary and secondary stress, and presence of 1s gap
Seidenberg & Elman	Syllable	Distributed — Binary	12 phonetic features
Altmann and Dienes	Syllable	Localist — Binary	
Shultz	Triplet	Localist* — Analog	*Localist for syllables
Sirois et al.	Syllable	Localist — Analog	
Shastri&Chang	Syllable	Distributed — Binary	6 phonetic features
Gasser&Colunga	Triplet	Localist* — Binary	*Also includes an *angle*
Domeney&Ramus	Syllable	Localist — Binary	

Table 6.3: Encoding used in the neural network models reviewed. In the scheme column: L stands for Localist, D for Distributed, B for binary, and A for analog.

for future work, as explained in desideratum 1.

Second, the fact that a neural network may show only some degree of generalization (by predicting a vector that is *close* to the 'correct' vector) begs the question of whether infants in the experiment are producing accurate generalizations. The analyzed empirical data is based on looking times, and therefore, there was no chance to observe if the discrimination between grammars was based on perfect generalization. It is therefore not clear whether we should expect models to produce perfect generalization or statistically significant responses between grammars. For this reason, I propose that models are evaluated at least on both aspects, as I reflect later in the desiderata (section 6.6, desideratum 6).

Finally, an aspect of the representation of the input that has not received attention is the treatment of time. Almost all the neural network models reviewed receive the input as discretized units, and update the weights of the model after each presentation (sometimes during a few timesteps, as in Shastri and Chang [1999]). The only exception is the spiking neural network model by Dominey and Ramus [2000], but even in this model we can find discrete syllable registers in its short-term memory. The use of discrete input has also forces the model to have a very unnatural representation of pauses, which are generally coded as one symbol –as if it were one more item in the vocabulary. For this reason, in the desiderata I suggest to investigate generalization over continuous input (desideratum 9).

6.5.2 Question 2: What is the learning process?

One of the most relevant and ambitious scientific questions behind the Marcus et al. experiment is to understand the nature of the learning mechanism that is operating during the experiment. Since we only observe the input stimuli and the behavioural output, the characterization of the unobserved procedure that relates them allows for multiple hypothesis.

In neural networks, the process of learning is commonly referred to as "associative learning", and it is characterized by responding to contingency relations in the data. Most neural networks are trained with some form of gradient descent; in the majority of cases, the algorithm used is backpropagation [Rumelhart et al., 1988]. Although the neurobiological plausibility of backpropagation was initially in question [Zipser and Andersen, 1988, Crick, 1989, Stork, 1989], later work argued that the learning procedure can be implemented also with biologically plausible bidirectional activation propagation [O'Reilly, 1996, 1998]; moreover, Xie and Seung [2003] prove that – under certain conditions– backpropagation is mathematically equivalent to Contrastive Hebbian learning, a process whose biological plausibility is widely agreed upon. All the neural network models reviewed implement some form of gradient descent, with the exception of the proposal of Dominey and Ramus [2000], which features a learning algorithm based on cortico-striatal circuits in the brain [Dominey, 1995], which researchers have interpreted as a form of Least Mean Squares [Lukoševičius, 2012].

In the rule-based model proposed by Frank and Tenenbaum [2011], the model learns through Bayesian inference over a predefined space of hypothesis. Therefore, this model does not offer an account of the process that induces the regularities in the first place. However, the authors clearly state that the model is proposed as an "ideal learner" rather, and thus it should not be interpreted as a model of human learning. Therefore, a cognitively realistic rule-based model that explains how learning takes place during the experiment is still lacking.

6.5.3 Question 3: Which generalization?

The speech streams we are concerned with are generated according to an ABA, AAB or ABB pattern, which involves relations between syllables. However the stimuli are also compatible with other rules, and – as discussed in section 6.5.1– regularities may also appear on other acoustic dimensions.

In order to illustrate this, table 6.4 shows some of the rules that describe the relations between syllables in the triplets. These can be as general as '*three consecutive syllables*' (equivalent to rule (a)), or they could operate over two of the syllables in the triplet. These basic rules can be composed with logical operators (*and, or, not*), such as '*(a) and (b)*'; for instance, if a learner is hypothesized to learn a rule like '*triplet containing an adjacent repetition*', this can be expressed as '*X=Y or Y=Z*'. As will be explained in next section, theories that postulate that rules are cognitively real need to disambiguate which rule is being learnt when rules are equivalent in their extension.

From the models have reviewed, only the Bayesian approach Frank and Tenenbaum [2011] explicitly tackles this question. In order to distinguish between otherwise equiprobable consistent hypotheses, this model incorporates a predefined rational principle that determines which hypothesis should be favoured; in this case, the *size principle*. In this way, this model offers a transparent way to test how different principles would be favoured by a probabilistic inference process, an aspect that is missing in the neural network models.

It must be noted that rational principles are not the only source of disambiguation to decide between competing rules. For instance, Endress et al. [2005] report experimental evidence showing that repetition-based grammars are easier to learn when the repetition takes place in the edge positions. This entails that other aspects –such as perceptual factors– can also impose saliency in certain dimensions of the stimuli, breaking the uniformity between otherwise equivalent rules. For this reason, I suggest in desiderata 2, 3 and 4 that alternative factors that influence why certain rules are favoured should be explored.

6.5.4 Question 4: What are the mental representations created?

Even if we agree on the generalization procedure and the question of which rule has been extracted, we still need to discuss how the induced rule is represented in the cognitive system.

Symbolic and non-symbolic models have different strengths and weaknesses. Rule-based models exploit the fact that symbolic rules can easily accommodate some of the most interesting properties of thought and language, such as systematicity, productivity and compositionality. The use of variables that are blind to the specific properties of their content ensures a rigorous description: as well as a binary output: rules are either consistent or inconsistent, never in between. However, this has the downside of endowing the models with little flexibility, and thus they are not robust to noise [Opitz and Hofmann, 2015]. In contrast, neural network models do not explicitly represent rules or variables, so relations are content-dependent (as well as context dependent). One of the advantages of these models is that they can naturally account for degraded instances or accidental gaps; therefore, exceptions can be handled without the need of additional mechanisms [Elman, 1999].

As researchers, we are used to employ formal languages for scientific descriptions, and thus it is natural to characterize stimuli with formal rules such as *XYZ such that X=Z*. And indeed, the behavior shown by the infants in Marcus et al. experiment can be described as following an identity rule that accounts for the familiarization grammar. However, the mental representations of infants in the experiment need not directly map with the components of such formal expression: the variables and the logical operators that relate them may or may not correspond directly to mental entities.

As argued in Pylyshyn [1991], it is common in science that a debate arises when the object of research involves a system that can be easily described with rules. The author outlines a topology for theories addressing those type of systems, according to which the ontological status of the theory can be seen as a point in a spectrum between two extremes. Theories may, on one extreme, postulate rules that only account for regularities in the behavior of the system. In this case, rules function only as a theoretical descriptive construct, but the theory is agnostic towards the representations and the principles *followed* by the system. In intermediate positions, theories postulate that *some* of its elements correspond to principles or properties materialized in the system. And on the other extreme, theories maintain that all its rules and representations are explicitly encoded in the system. In the latter case, the elements in the canonical expression of a rule (including its symbols and the relations between them) correspond to certain properties in the system. Thus, details such as the total number of rules and their precise definition (e.g. whether they are based on identity or difference, even if their scope is identical) become relevant for a theory that posits that these rules are materialized in the cognitive system.

This characterization of theories can be easily related to Marr's levels of analysis [Marr, 1982]. Marr suggests a topology for computational models of cognition, such that: i) *computational level* theories are concerned only with characterizing the problem and the solution, ii) *algorithmic* or *processing level* models propose a mechanistic account of the process and the representations in the cognitive system, and iii) *implementational level* approaches explain how the process and representations are physically instantiated. Thus, according to this characterization, computational-level models may employ high-level descriptions that may not translate into actual representations in lower levels.

With this taxonomy at hand, it is straightforward to see where the described models stand on such issue. The rule-based model in Frank and Tenenbaum [2011] is explicitly stated at Marr's computational level; therefore, the fact that it employs symbolic rules is not to be taken as a representational claim. On the other hand, neural network models are implementational-level accounts of how processes and representations may be realized. Thus, (most of) the reviewed neural network models propose that symbolic representations are not cognitively real, and that the emergent behaviour observed in the experiment can emerge even when symbolic representations are not employed.

6.5.5 Conclusion of the analysis

This analysis has allowed us to closely examine how this collection of models has helped in advancing our knowledge on the main research questions. But surprisingly, in spite of the relative simplicity of the experiment and the vast number of models, the state of our knowledge appears rather incomplete when we analyze the questions at this level of detail. For this reason, I have compiled an agenda of the issues that require more attention.

6.6 An agenda for Rule Learning

I now reflect on how the unresolved issues identified in our analysis could be materialized in future studies, and I briefly mention some of the most recent work that is starting to address my desiderata.

1. DESIDERATUM. *Investigate which features should be encoded in the input representation, and quantify the overlap of features needed for generalization to occur.*

To begin with, not much attention has been devoted to how the perception of the input can affect generalization. The syllables in the original experiment are chosen to minimize phonetic overlap, but as pointed out by McClelland and Plaut [1999], other acoustic cues may exhibit regularities. Additionally, similar experiments involving a different set of syllables show null results (Geambasu&Levelt, p.c.), suggesting that low-level cues are relevant.

As mentioned before (section 6.5.1), the amount of overlap in the representation of input vectors in neural network models influences the prediction of novel items. Thus, more research is needed to quantify the amount of overlap required to reproduce the empirical findings, and specially, which features should be encoded in the vectors (and therefore, which perceptual dimensions guide generalization).

2. DESIDERATUM. *Investigate perceptual biases.*

The second issue that can be observed is that, in all models, the perceptual units (generally syllables) are treated equally, regardless of the position they appear in. However, as mentioned before, experimental work shows that syllables that appear in the edge of sequences are more salient to humans, to the extent that some rules are not learnt if the regularity appears in middle positions [Endress et al., 2005].

The reviewed models do not explicitly incorporate any such biases. Although a case could be made for neural networks being able to *learn* those biases, this would only occur when saliency facilitates the task. Thus, I suggest that future efforts should be directed to investigate which perceptual biases facilitate or hinder generalization.

3. DESIDERATUM. *Investigate the role of prior experience.*

Most of the models reviewed are used as a *tabula rasa*: they are initialized with some independent method (e.g. randomly sampled weights in a neural network) and then trained exclusively on the familiarization data. However, for a randomly initialized model, it is unlikely that a short exposure to the familiarization stimuli suffices to reproduce the experiments. If, instead, the initial state of the models incorporates relevant prior knowledge, the learning procedure may converge more easily to the generalizing solution that infants seem to learn.

This is the idea behind the models proposed by Altmann [2002] and Seidenberg and Elman [1999a], but our analysis concluded that these models are not convincing explanations of the empirical phenomenon. Moreover, the use of pretraining procedures would result more explanatory if the they allow to pinpoint which aspects of the prior knowledge are the ones influencing generalization to novel items.

4. DESIDERATUM. *Model the coexistence and competition of rules.*

It is very unlikely that a perceived flow of information can be described with one single rule, but rather, an input stream is likely to incorporate regularities between features in different dimensions. Thus, a model of generalization should explain how the detected regularities coexist, that is, how they are represented and whether they interfere between each other during the course of perception and learning.

In reviewing the existing models of generalization of simple grammars, I have observed that the question of "which generalization" has been widely neglected. The Bayesian model by Frank et al. has initiated that an approach to that question by proposing the size principle as a rational criteria to predict a preference for some rules over others (see also Chater and Vitányi [2003] for an more general argument for *simplicity* as the disambiguating rational principle). However, I argue that factors other than rational principles of parsimony may also play a role.

Some of the desiderata outlined above suggests factors that may influence the preference for some rules, such as perceptual factors like edge saliency, or the role of prior experience in shaping the perceived contingencies (or in providing the right pressure for finding generalising solutions). External factors may also play a role; for instance, the contextual information.

As an example, consider a math student who is asked to generalize after seeing numbers 30 and 40: we would expect her to be more likely to generalize to number 50 than to 41, inducing a mathematical rule such as 'multiple of 10'. But a medicine student who is studying dangerous cholesterol levels, after observing that 30 and 41 are a dangerous levels, would prefer to generalize to 41 rather than 50, due to proximity. Given the same input, and the task of generalizing to novel stimuli, the context provides extraneous cues that have an effect on the favoured generalizations (example adapted from [Tenenbaum and Griffiths, 2001]).

To sum up, I suggest that future models attempt to explain how rules coexist and compete, and aim to reveal which factors determine which rules are preferred.

5. DESIDERATUM. *Incorporate independently motivated pressures for learning generalizing solutions.*

The hypothesis space in neural networks often contains multiple local optima, and thus the learning procedure has high risk of getting stuck in one of those local optima. This entails that, in practice, there exist multiple solutions that may be found by a neural network model. While these solutions can be sufficient to account for the training data, they may not be *generalising* solutions that can be transferred to the test stimuli.

This can be seen as a form of *overfitting*. Since neural networks have many degrees of freedom, they can easily find one of the non-generalizing solutions. In order to push a neural network model to find a generalizing solution, an additional source of pressure is needed. Thus, I believe an important avenue of future research is to investigate how to incorporate independently motivated biases to find generalizing solutions.

6. DESIDERATUM. *Find a consensus in evaluation criteria for models.*

One of the most pressing issues that I have discovered in this review is that there is ample disagreement in the formulation of the objective to optimize and, specially, on how to evaluate the outcome of a model.

As I observed before, reflecting on the evaluation reveals an important issue about the empirical data: while the results exhibit statistical significance for the attention that infants show between different grammars, there is no evidence that infants perform accurate generalization (i.e. correct prediction of the last syllable in a triplet). It is certainly ambitious to establish the reality for the subjects in the experiment, since we do not know of experimental procedures that allow us to investigate what exactly have infants learnt. Progress in this direction would definitely facilitate that modellers settle for one particular evaluation method and compare the different model proposals in a systematic manner.

For lack of better knowledge, I suggest that models are evaluated on both criteria: i) whether they exhibit statistical significance for grammar discrimination, and ii) whether the model accurately generalizes to new items.

7. DESIDERATUM. *Bridge the gap between levels of analysis: investigate how neural networks perform apparently symbolic computations.*

As discussed before, one of the most debated issues is the ontological status of symbolic rules and variables. Even though the neural network models reviewed have arguably shown some success in reproducing the empirical findings, the relation between the symbolic-like behaviour and its actual realization in the model is not completely clear: neural networks are assumed to perform symbolic-like computations implicitly, but how exactly this is done remains elusive. Thus, in order to understand how non-symbolic systems perform apparently symbolic computations, we should aim to investigate the internal representations and strategies employed by neural networks.

This issue can be regarded as finding the relation between computational level explanations (for which rules and symbols are well suited) and the corresponding implementational realization. Thus, another way to frame this problem is in trying to bridge the gap between the computational level approaches (such as those offered by Bayesian models) and implementational models.

8. DESIDERATUM. *Models should learn spontaneously from brief/limited exposure.*

Neural network models need to iterate over the data in order to converge to a good parameter setting. This is aggravated in the case of small training datasets (as would be the case for the Marcus et al. stimuli), since less data entails that more epochs are required for convergence. Unfortunately, this does not reproduce the training set up used in the infants experiment, in which the familiarization stimuli is presented only once. Although it could be argued that the iterations of the stimuli reflected the availability of the data in short term memory (such as the phonological loop), this does not allow us to distinguish between tasks that can be learnt spontaneously and those that require a longer exposure or even a developmental trajectory.

For this reason, I think that achieving spontaneous learning with neural network models is an ambitious but relevant project. Recent advances on neural network architectures involving augmented memory (e.g. the Neural Turing Machines [Graves et al., 2014]) are capable of fast memorization and retrieval of input stimuli, and thus they offer a promising avenue for learning from short-exposures (e.g. successes in one-shot learning have been reported in Santoro et al. [2016]).

9. DESIDERATUM. *Investigate the effect of the continuous nature of speech in generalization.*

An aspect that has been widely neglected in the reviewed models is to represent the continuous nature of speech. Some of the models operate over all the data at once (concretely, the Bayesian model), while the neural networks process the data in an online fashion, either over triplets, syllables or phonemes. But even in this case, the stream is pre-segmented, and the models are updated synchronously in discrete timesteps – although, as argued in section 6.3.3, the model by Dominey and Ramus [2000] offers a more realistic treatment of time, but it does not succeed in modelling the experiment without pre-segmenting the input syllables and accessing its storage in a symbolic fashion.

By simplifying the representation of time, some aspects of auditory processing can be neglected. For instance, the speed at which an auditory speech stream is played may have an effect on the structural dependencies that learners can extract from it, due to the temporal proximity of the items involved. This phenomenon cannot be modelled with discrete neural networks, since there is no manipulation that can account for the speed of presentation of the syllables. It remains an open question whether a more realistic treatment of time would also bring new insights to the question of generalization.

6.7 Conclusions

The study by Marcus and colleagues has been very influential in the field, thanks to showing generalization abilities in 7 m.o. infants that had not been previously attested. It is fair to point out that these results have proved to be difficult to replicate (see footnote 1 in Gerken [2006] as an example; also Geambasu&Levelt, p.c.); likewise, other generalization experiments [Gomez and Gerken, 1999] show behavior in the opposite direction, with infants looking significantly longer to consistent rather than inconsistent test items. Thus, unraveling under which conditions the Marcus et al. study can be replicated requires further investigation, as well as a methodology to compare the outcomes of a group of related studies (e.g. see methodological proposal for meta-analysis of experimental data in Tsuji et al. [2014]). Nevertheless, even if I conclude that the experiment can only be replicated under very specific conditions, its design has undeniably been a very fruitful for posing concrete questions about the nature of generalization.

In this work, I have contrasted different modelling traditions. Even though some approaches may appear irreconcilable at first glance, it actually seems that neural networks and Bayesian approaches are somewhat complementary in their contribution to understanding generalization. The former offer a theory of how humans discover the relevant regularities in the input, while the latter provides a transparent method for deciding between alternative hypothesis. In neural networks, it is complicated to predict beforehand which of the competing hypothesis would be learnt, and thus, as explained in the previous section, the choices for the represented dimensions of the input will have an impact on what is learnt. In contrast, in Bayesian models we can test rational principles to distinguish between possible generalizations, but those models require a pre-specified hypothesis space and a cognitive theory of how the cognitive process takes place.

Overall, our study shows that, in spite of the many questions that can be raised from the Marcus et al. study and the large number of modelling contributions, most of the discussion has been centered around the question of the ontological status of rules and symbols. Although I agree that one of the most intriguing issues in cognitive science is to discover whether rule-like behaviour requires symbolic operations, I expect that our review brings back attention to other aspects of the problem that have received less attention, such as the impact of perceptual factors and the question of which rule is preferred among competing consistent rules. Hopefully, future experimental work and modelling efforts in these directions would help unravelling the underpinnings of generalization.

	ABA	ABB	AAB	Consistent
(a)	XYZ	XYZ	XYZ	
(b)	$XYZ : CVCVCV(CV : Consonant - Vowel)$	$XYZ : CVCVCV$	$XYZ : CVCVCV$	
(c)	$XYZ : X \neq Y$	$XYZ : X \neq Y$	$XYZ : X = Y$	
(d)	$XYZ : Y \neq Z$	$XYZ : Y = Z$	$XYZ : Y \neq Z$	1,2,3
(e)	$XYZ : X = Z$	$XYZ : X \neq Z$	$XYZ : X \neq Z$	1,2
(f)	presence of repetition	presence of repetition	presence of repetition	
(g)	presence of nonadjacent repetition	presence of adjacent repetition	presence of adjacent repetition	1,2
(h)	voiced-unvoiced-voiced	voiced-unvoiced-unvoiced	voiced-voiced-unvoiced	1

Table 6.4: Summary of some of the rules that the learners in Marcus et al. may extract. Column "consistent" indicates whether the rules suffice to explain results in experiments 1, 2 and/or 3. These rules could further be composed with *and*, *or* and *not* operators (e.g. the composition $(c)AND(d)$ suffices to explain the three experiments).

Chapter 7

Pre-Wiring and *Pre-Training*: What does a neural network need to learn truly general identity rules?

In the previous chapter I reviewed a range of models of generalization in ALL, and reflected on the next steps needed to increase our knowledge on this mechanism. I now present in this chapter[1] a new model that results from addressing many of the proposed desiderata: in particular, investigating the role of prior experience (desideratum 3) and learning biases that facilitate generalization (desideratum 5), but also understanding the relation between neural and apparently symbolic computations. I also make progress on desideratum 6 by proposing two different complementary methods of evaluation for models of generalization. Aspects of desiderata 4 and 8 are briefly explored as well.

7.1 Introduction

Accounting for how humans learn abstract patterns, represent them and apply them to novel instances is the central challenge for cognitive science and linguistics. In natural languages there is an abundance of such phenomena, and as a result linguistics has been one of the main battlegrounds for debates between proponents of symbolic and connectionists accounts of cognition. One of the most heated debates was concerned with accounting for the regular and irregular forms of the English past tense. Rumelhart and McClelland [1986] proposed a connectionist model that allegedly accounted for the regular and irregular forms of the past tense. However, this model was fiercely criticized by Steven Pinker and colleagues [Pinker and Prince, 1988, Pinker, 2015], who held that rules are essential to account for regular forms, while irregular forms are stored in the lexicon (the 'Words-and-Rules' theory).

[1] The content of this chapter is based on the following publication:
Alhama and Zuidema [2017a]. Pre-Wiring and Pre-Training: What does a neural network need to learn truly general identity rules? [*Under review.*]

As seen in the previous chapter, a similar debate emerged with the publication of Marcus et al. [1999], this time centered on experimental results in Artificial Grammar Learning. The authors showed that 7 month old infants generalize to novel instances of simple ABA, ABB or AAB patterns after a short familiarization. Crucially, this outcome could not be reproduced by a Simple Recurrent Network (SRN) [Elman, 1990], a result that was interpreted by the authors as evidence in favour of a symbol-manipulating system:

> Such networks can simulate knowledge of grammatical rules only by being trained on all items to which they apply; consequently, such mechanisms cannot account for how humans generalize rules to new items that do not overlap with the items that appear in training. [Marcus et al., 1999, p. 79]

This claim triggered many replies, some of which proposed variations of the original model. However, in this debate the issues of whether neural networks are capable at all of *representing* general rules, of whether backpropagation is capable of *finding* these general rules from an arbitrary initial state or only from an appropriately chosen initial state are sometimes conflated. The latter issue – what initial state does a neural network model need to have success in the experiment – has, in my view, not received enough attention (but see Seidenberg and Elman [1999a], Altmann [2002]). This will be therefore the focus of this chapter, in which I explore two directions. First, I ask which initial values of the connection weights could encourage generalization while remaining cognitively plausible (*pre-wiring*); second, I investigate the role of previous experience in creating an initial state in the network that would facilitate generalization (*pre-training*). I employ a prewiring and a pretraining technique in an Echo State Network (ESN) [Jaeger, 2001], and show that only when combining both techniques the ESN is able to accurately generalize to novel items.

7.2 Background

7.2.1 Empirical Data

In this chapter, I focus on modelling the empirical data presented in Marcus et al. [1999]. Since this study has been extensively discussed in the previous chapter, the reader can refer to § 6.2 for more details. For convenience, I also briefly summarize the study here.

Marcus and colleagues investigate the generalization abilities of 7 month old infants by conducting three Artificial Language Learning experiments. In their first experiment, the participants are familiarized to syllable triplets that follow a certain grammar: ABA for a randomly assigned group of infants, and ABB for the other. The stimuli contain 16 different triplets, each repeated 3 times. Those triplets are arranged in a 2-min. auditory speech stream, such that syllables are separated by a pause of 250 ms, and triplets of syllables are separated by 1s.

After the familiarization, the infants participate in a test phase, in which their looking times (to the speaker device that plays the stimuli) are recorded. The speaker plays a randomized set of triplets from both grammars, in order to see if infants can discriminate between them. Crucially, the test triplets contain syllables that were not used in the familiarization stimuli.

The results show a statistically significant difference between mean looking times to consistent and inconsistent grammars in both group of infants. The authors then conclude that infants can discriminate among ABA and ABB grammars.

An additional experiment was performed, in this case using AAB vs. ABB grammars, in order to determine whether the rule learnt before was simply the presence or absence of an immediate repetition. Infants also showed significantly different responses in this experiment.

In the light of these results, the authors concluded that: (i) 7 m.o. infants can extract grammar-like rules, (ii) they can do it not based solely on statistical information (as would be evidenced from the additional controls in experiment 2, and (iii) the extracted rule is not merely the presence or absence of an immediate repetition.

7.2.2 Generalization and Neural Networks

Marcus [1998] argues that certain types of generalizations are unattainable for certain types of neural networks: concretely, those that lie *outside the training space*. The author defines *training space* as the combination of all feature values that network has witnessed during training. If there exist feature values that have never appeared during training, any item displaying that feature value lies outside the training space. For neural networks that are trained with the backpropagation algorithm, generalization to items outside the training space is, according to the author, extremely unlikely to occur due to what he calls *training independence*, which stands for the fact that the algorithm updates the weights of nodes independently of the activations of other nodes in the same layer.

In Marcus et al. [1999], the authors provide empirical evidence in support of this idea, by simulating the presented experiment in a Simple Recurrent Network (SRN) [Elman, 1990], a neural network architecture that incorporates an additional context layer that maintains an exact copy of the hidden layer and presents it to the network in the subsequent timestep, providing the model with memory in this way. The SRN is trained to predict the next syllable in the familiarization stimuli, and then tested on its ability to predict the final syllable of test items consistent with the familiarization grammar. This model failed to produce the correct predictions, confirming the hypothesis of the researchers.

Some following publications proposed to change the encoding of the input [Christiansen and Curtin, 1999, Christiansen et al., 2000, Eimas, 1999, Dienes et al., 1999, Altmann and Dienes, 1999, McClelland and Plaut, 1999], the task [Seidenberg and Elman, 1999a,b], the neural network architecture [Shultz, 1999, Sirois et al., 2000, Shultz, 2001], or – relevant to this work — incorporating some form of pre-training

[Seidenberg and Elman, 1999a, Altmann, 2002]. Many of these models were subject of criticism by Marcus [Marcus, 1999a,b,c,d], who argued that the models either involved some form of symbolic manipulation or did not adequately represent the experiment. About the model of Altmann [2002], which involves pre-training similar to the regime explored in section 7.6, Marcus [1999e] points out, without giving any details, that, even if the model distinguishes grammatical from ungrammatical stimuli to some degree, it is unclear whether the model can actually learn the underlying general rule or discovers some heuristic that weakly correlates with it. In my work, I employ a neural network architecture that was not previously explored for this task (an Echo State Network, a type of Reservoir Computing network), and report additional performance measures that tell us more about how general the learned rules are.

7.3 Simulations with a Simple Recurrent Network

Before presenting the simulations with the ESN model, I replicate the original simulations. I implement a Simple Recurrent Network as described in Elman [1990], and train it to predict the next syllable in the input. As in Marcus et al., I use distributional encoding of phonetic features (based on Plunkett and Marchman [1993]). But unlike the original simulations, I do not encode the pause between triplets as an additional symbol; instead, I do not update the weights in the network when it predicts the first syllable of the next triplet (I do this to make my baseline results maximally comparable with the simulations presented in the next sections).

In order to remain close to the test used in the experiments with infants, I test the network on both consistent and inconsistent sequences. I take the predicted vector for the third syllable of each triplet, and I find the closest vector that corresponds to one of the seen syllables (both from training and from test). I then evaluate whether the accuracy for consistent and inconsistent triplets is significantly different (for 16 runs of the model, equivalent to the number of infants in the experiment).

The test set used in the original experiments, as can be seen in Table 1, is based solely in two triplet types of each grammar. For this reason, I also evaluate my model with an extended test set that contains 5 additional random novel syllables of each type (A and B), consisting therefore of 25 test triplets.

I try 160 parameter settings for each familiarization grammar, varying the hyperparameters of the model: the size of the hidden layer, the learning rate and the number of epochs [2] . Figure 7.1 shows the proportion of these runs that yield a significant difference in the predictions for the two classes of test items (those that are consistent with the grammar used in training and those which are not). For the responses that are significantly different, I separate those for which the neural network responds better to the consistent grammar (in white) from those in which the inconsistent grammar is favoured (in grey).

[2] I found that the values of the three hyperparameters had a significant effect on the accuracy of the predicted syllables in the test.

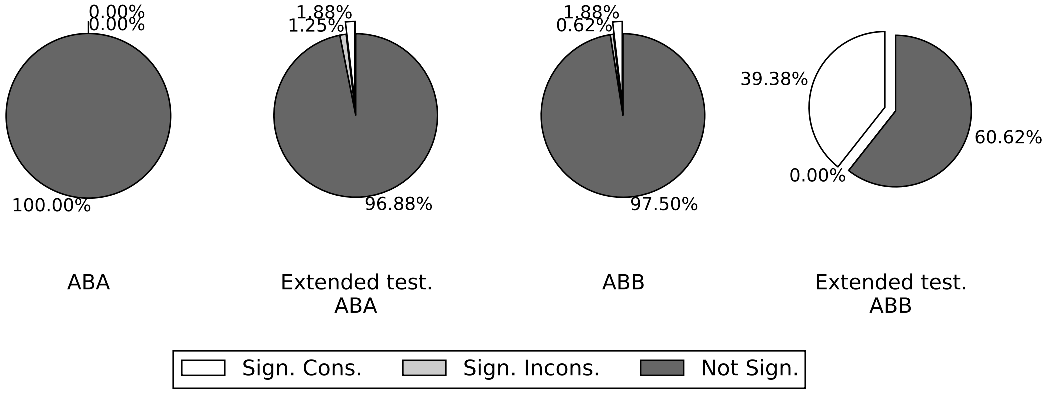


Figure 7.1: Proportion of parameter settings that yield significant (white), non-significant (dark grey) and inconsistently significant (light grey) differences in the responses between grammars, for simulations with an SRN.

As shown in the graphic, most of the simulations yield non-significant response differences between grammars, in spite of a notable proportion of significant responses in the ABB condition for the extended test, possibly due to the fact that immediate repetitions are easier to learn[3]. I therefore confirm that the Simple Recurrent Network does not reproduce the empirical findings.

7.4 Simulations with an Echo State Network

Recurrent Neural Networks, such as the SRN, can be seen as implementing memory: through gradual changes in synaptic connections, the network learns to exploit temporal regularities for the function it is trained on. An alternative way to learn time-dependent relations is that offered by Reservoir Computing (RC) approaches, such as the Liquid State Machine [Maass et al., 2002] and the model adopted here, the Echo State Network (ESN) [Jaeger, 2001, Frank and Čerňanský, 2008]. In RC models, the weights in the hidden layer (which is dubbed "reservoir") remain untrained, but – if satisfying certain constraints (the so-called "Echo State Property", which depends on the scaling of the weights in the reservoir based on the spectral radius parameter) – the dynamics exhibited by the reservoir "echo" the input sequence: some memory of the input lingers on for some time in the recurrent connections. In other words, the state of the reservoir depends on the fading history of the input; after a long enough input, the initial state does not determine the final states of the network.

The formalization of the ESN model is as follows. For an input u at time t, the activation x of the nodes in the reservoir is defined as:

[3] This was also observed in the SRN model in Altmann [2002].

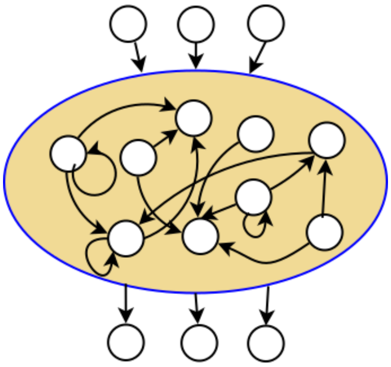


Figure 7.2: The Echo State Network.

$$x(t) = f(W^{in} \cdot u(t) + W^{res} \cdot x(t-1)) \tag{7.1}$$

where W^{in} are the input weights, W^{res} are the internal weights of the reservoir, and f is a non-linear function, generally *tanh*.

The activation of the output is defined as:

$$y(t) = f^{out}(W^{out} \cdot x(t)) \tag{7.2}$$

where W^{out} are the weights that connect the reservoir with the output nodes, and f^{out} is a function, which might be different from the function applied to the reservoir; in fact, it often consists of a simple identity function.

I implement a basic ESN with tanh binary neurons, and I follow the same procedure described in section 7.3 to train the network with backpropagation[4]. I try 200 parameter settings for each familiarization grammar, varying the hyperparameters of the model: the number of nodes in the reservoir, the input scaling, the spectral radius, learning rate and epochs.[5] Figure 7.3 shows the proportion of these runs that yield a significant difference in the predictions.

As can be seen, the results based on the Marcus et al. test set differ greatly from those in the extended test. This confirms my intuition that the amount of test items is crucial for the evaluation. For this reason, I base the analysis of the behaviour of the model in the extended test; however, it is important to notice that the amount of test items could have also played a role in the actual experiments with infants (see also section 7.8).

The plots of the extended test condition clearly show an assymetry between the grammars: more than half of the parameter settings yield significant responses for the ABB, while in the case of ABA, less than a quarter of the simulations are significant, and most of them are actually favouring the inconsistent grammar, which is precisely

[4] I have also run simulations with Ridge Regression, with similar results.

[5] I found that the values of the input scaling and the learning rate had a significant effect on the accuracy of the predicted syllables in the test.

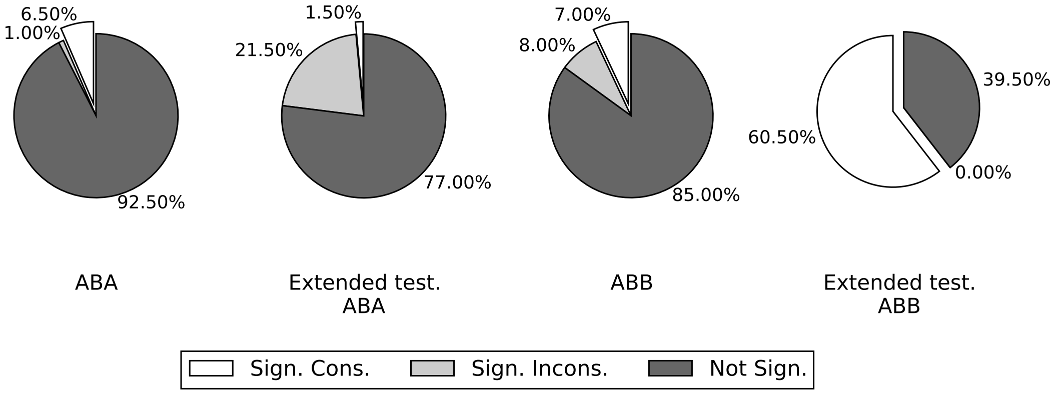


Figure 7.3: Proportion of parameter settings that yield significant (white), non-significant (dark grey) and inconsistently significant (light grey) responses, for the simulation with the basic ESN.

ABB. As mentioned before, the reason for this assymetry is probably due to the fact that immediate repetitions are easier to learn, since they are less affected from the decay of the activation function; for now, it suffices to say that the behaviour of the model towards ABA does not suggest that it could be a potential explanation for the experimental results.

7.5 *Pre-Wiring*: Delay Line Memory

In order to succeed in the prediction task, the model must predict a syllable that is identical to one presented before. In the previous simulations, I relied on the memory that ESNs offer through the recurrent connections and the Echo Property [Jaeger, 2002]. However, there exist several computational alternatives that brains may use to implement memory [Chaudhuri and Fiete, 2016]. I now explore one such model of memory: a delay line.

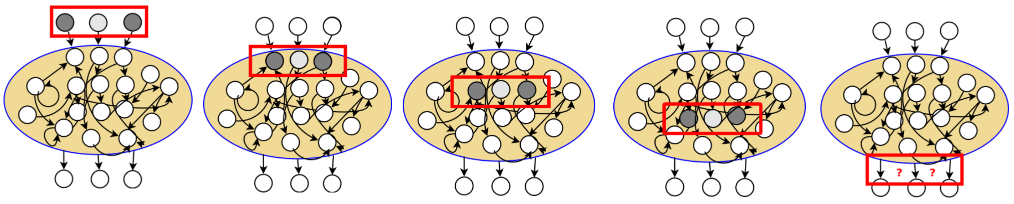


Figure 7.4: Depiction of five timesteps in the DeLi-ESN. The highlighted nodes show the activations in the delay line (the activation of the rest of the nodes is not illustrated).

Computationally, a delay line is a mechanism that ensures the preservation of the input by propagating it in a path ("line") that implements a delay. In the brain, delay lines have been proposed as part of the sound localization system [Jeffress, 1948], and they have been identified through intracellular recordings in the barn owl brain [Carr and Konishi, 1988]. In a neural network, a delay line is naturally implemented by organizing a subnetwork of neurons in layers, with "copy connections" (with weights 1 between corresponding nodes and 0 everywhere else) connecting each layer to the next. In this way, the information is kept in the network for as many timesteps as dedicated layers in the network (see figure 7.4).

I implement a delay line memory in the ESN (creating thus a new model that I call *DeLi-ESN*) by imposing this layer structure in the reservoir. I run 1200 combinations of parameter settings with the DeLi-ESN, including also a parameter that establishes some amount of noise to add to the weights of the reservoir. In this way, some models contain a less strict delay line; the greater the noise, the closer the model is to the original ESN.

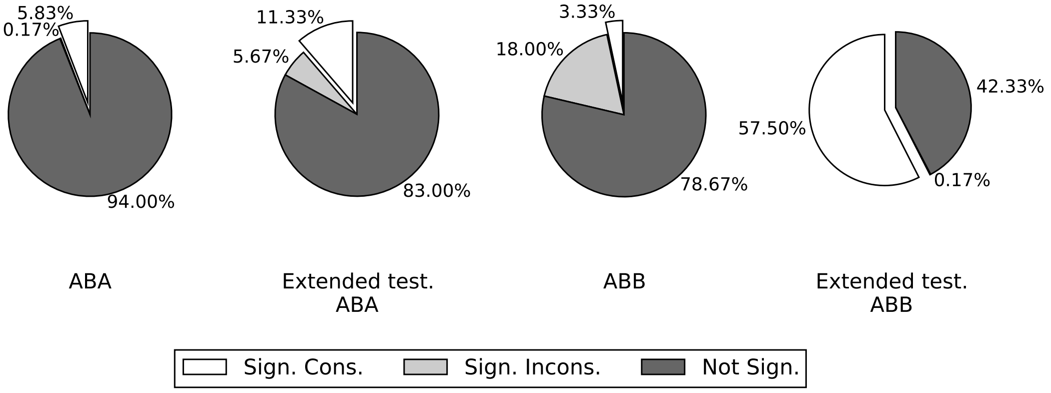


Figure 7.5: Proportion of parameter settings that yield significant, non-significant and inconsistently significant responses in the tests, for the simulation with DeLi-ESN.

The results, illustrated in Figure 7.5, show an increased number of significant responses (in favour of the consistent grammar) for the extended test of ABA familiarization: the addition of the delay line memory indeed helps in the detection of the identity relation. But in spite of the positive effect of the delay line, we need to ask ourselves to what extent these results are satisfactory. The pie plots show the likelihood of obtaining with my model the results that Marcus et al. found in their experiments, and in order to do so, I use the same measure of success (i.e. whether the responses for each grammar are significantly different). However, the models hardly ever produce the correct prediction[6]. For this reason, in the next section, I adopt a stricter measure

[6] I find that the values of the input scaling, learning rate, spectral radius, reservoir size, and reservoir noise each have a significant effect on the accuracy of the predicted syllables in the test, although exact

of success. I discuss this issue further in section 7.8.

7.6 *Pre-Training*: Incremental-Novelty Exposure

The infants that participated in the original experiment had surely been exposed to human speech before the actual experiment; however, in most computational simulations this fact is obviated. I hypothesize that prior perceptual experience could have triggered a bias for learning abstract solutions: since the environment is variable, infants may have adapted their induction mechanism to account for novel instances. I now propose a method to pre-train a neural network that aims to incorporate this intuition.

In this training regime —which I call Incremental-Novelty Exposure, or INE for short— I iteratively train and test my model; so for a certain number of iterations *i*, the model is trained and tested *i* times, with the parameters learnt in one iteration being the initial state of the next iteration. The test remains constant in each of these iterations; however, the training data is slightly modified from one iteration to the next. The first training set is designed according to the Marcus et al. stimuli: 4 syllables of type A and 4 syllables of type B are combined according to the pattern of the familiarization grammar (ABA or ABB). In the second iteration, one syllable of type A and one of type B are deleted (that is, all the triplets involving those syllables are removed from the training set), and a new syllable of type A and one of type B are incorporated, such that new triplets of the familiarization pattern are generated with the new syllables (combined as well with the already-present syllables). Therefore, for each training iteration, the model is exposed to a similar training set as the previous iteration, but there is a small amount of novelty. This procedure is illustrated in figure 7.6.

I simulate 600 different hyperparameter configurations, varying the reservoir size, noise in the delay line, input scaling, spectral radius, learning rate, and epochs. Figure 7.7 illustrates how the mean accuracy evolves at each stage of the INE procedure of one representative run. As it can be seen in the graphs, the accuracy is really low in the beginning (corresponding to a simulation without pre-training) but, with more iterations –and thus with more novel items incorporated in the training set–, the model becomes better, presumably by finding a more general solution.

In order to test that these results are robust, I compute the mean over the accuracy for the last quarter of the tests (in this case, the last 25 tests, corresponding to the rightmost curve in the graph), for a few runs. The results are fairly similar in each run, as can be seen in figure 7.8a. Thus, the combination of the DeLi-ESN and the INE drastically boosts the generalization capabilities of the ESN; however, we should identify what is the contribution of the DeLi-ESN. Figure 7.8b shows the mean accuracy (again, for the last quarter of the regime), for 16 runs of the basic ESN in the INE regime. The effect of the delay line memory is clear: when removed, the accuracy is close to 0 for ABA, and mostly around 20% for ABB.

prediction accuracy remains low (rarely above 12% for ABA and above 20% for ABB familiarization) even for the best combination of parameters.

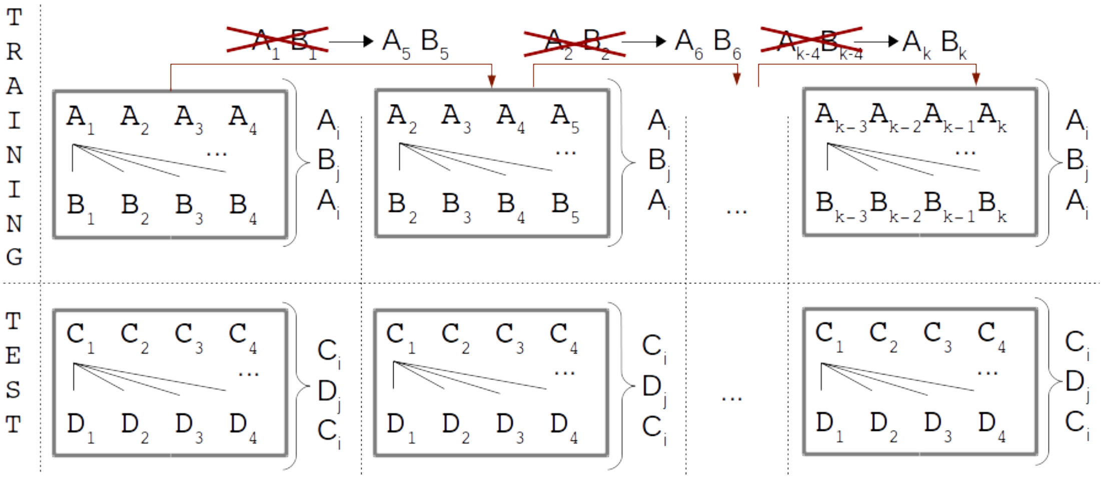


Figure 7.6: Depiction of the Incremental-Novelty Exposure. Barred syllables are removed from the training set after the training of the corresponding iteration has finished (so they do not remain in the training set of the next iteration).

7.7 Discussion: Relation with other neural network techniques

The task defined by Marcus and colleagues is simple, but the difficulty it presents for neural networks has been taken to reveal a fundamental limitation of such networks in real world tasks. I now therefore briefly discuss how my solution to these difficulties might relate to other methods in artificial intelligence and deep learning.

7.7.1 Relation with Recurrent Neural Networks

My model differs from the SRN model, and from Recurrent Neural Network (RNN) more generally, in its adoption of the *reservoir computing* approach. Concretely this means the weights from the input layer to the hidden layer, and the recurrent weights from the hidden layer to itself, are never updated. Thus, learning only occurs in the output layer.

This has important consequences for how each model implements a memory. In RNNs, input from previous time steps may continue to play a role through the recurrent connections. Since those connections are updated during training, the model must *learn* what to memorize and how to combine it with novel input. In contrast, in reservoir computing models, the reservoir is designed, by enforcing the echo property, to have informations from past inputs (and initialization) linger around in the recurrent connections (and this information is asymptotically 'washed out' [Jaeger, 2007]). Hence, the architecture does not need to learn to memorize information, but rather may focus on how to use both the memorized and current input to solve a task.

There is an interesting connection between the echo state property in reservoir com-

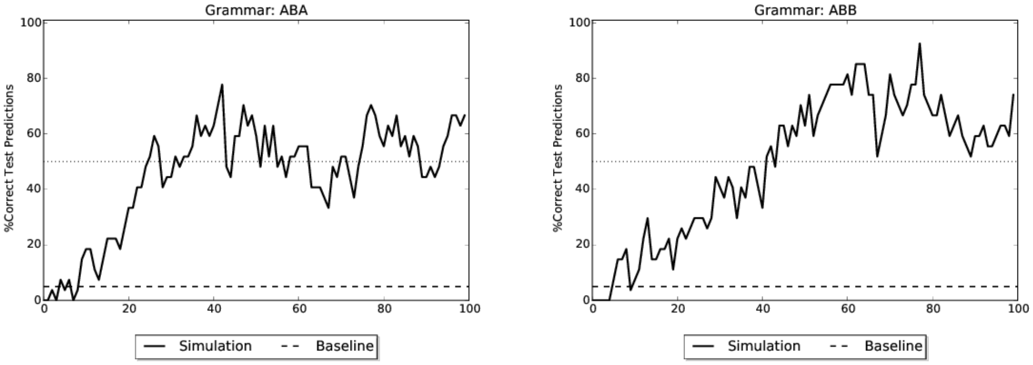


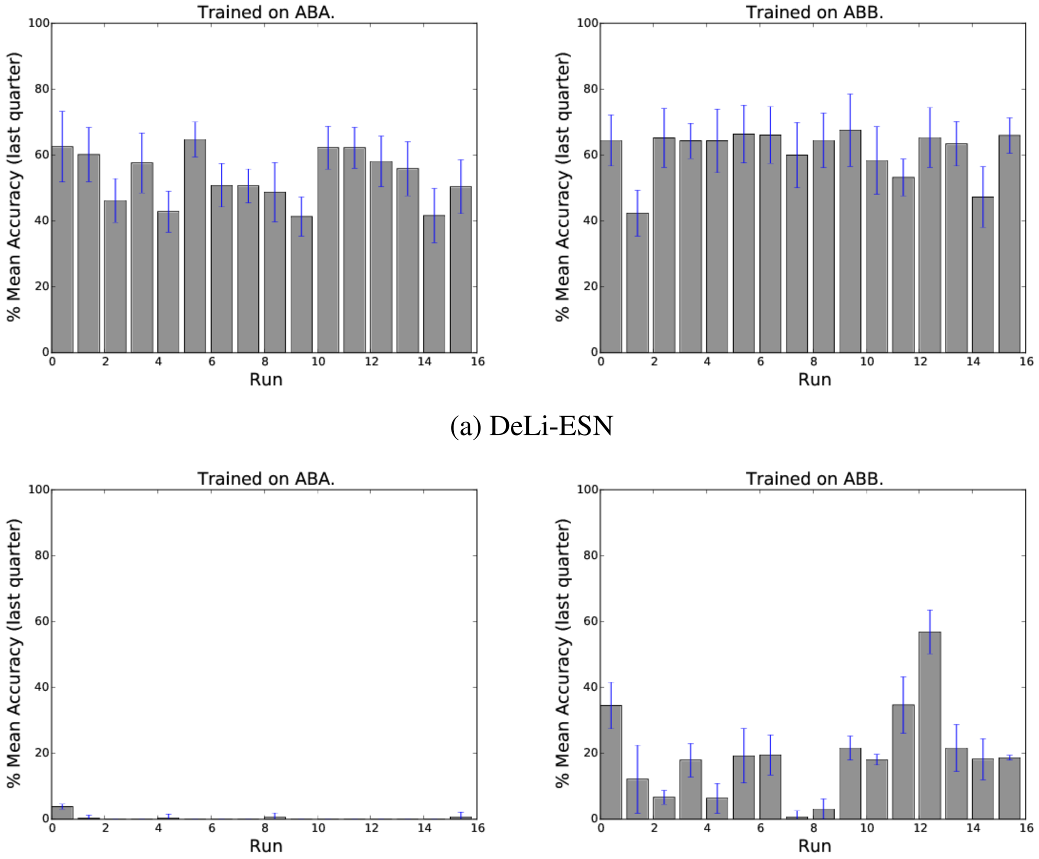
Figure 7.7: Performance over 100 iterations with Incremental Novelty Exposure, for one representative run in the ABA familiarization condition (left) and one in the ABB condition (right).

puting and the now popular ‘orthogonal initialization’ for recurrent networks. In echo state networks, the echo state property is often achieved by ensuring, as I do in here, that the spectral radius (largest eigen value) of the recurrent weight matrix has a value just below 1.0 (although this is neither a sufficient nor a necessary condition for the echo state property to hold; see Jaeger 2007). In ‘orthogonal initialization’ all eigenvalues of that matrix have an absolute value of 1; the largest eigenvalues will thus initially be 1.0, before backpropagation gradually changes the weights and eigen values.

7.7.2 Relation with standard Echo State Network

There are two main changes in my model that deviate from the standard ESN. First of all, by *pre-wiring* the reservoir, I depart from the original implementation, in which the reservoir is randomly initialized (although constrained to satisfy the echo state property). This has an effect on the type of memory the model has access to. In standard ESN’s, the echo state property ensures that the information of previous input lingers around, but not that this information remains represented in the same way: any (sub)vector representing the information of interest at time t might be inverted, rotated or projected into another subspace at any time t. Instead, the delay line in my model keeps information in the same representation (and projected to a dedicated subspace for each time delay).

Another aspect that deviates from common implementations of ESN is that I train through gradient descent, while generally, ESNs are trained with Ridge Linear Regression. The latter is a simple and fast solution, works over a convex error function, and converges to the global optimum. However, for the ’Incremental Novelty Exposure’ in section 7.6, the global optimum for the current training set is not what is looked for. Key to the success of that approach is that previous training phases still have an



(a) DeLi-ESN

(b) Basic ESN

Figure 7.8: Mean accuracy for the tests in the last quarter simulation in the INE simulation, for DeLi-ESN (a) and basic ESN (b).

effect on the current. With my use of stochastic gradient descent, my model explores multiple local optima, and thus has a chance to find a solution that generalizes beyond the current datasets.

7.7.3 *Pre-Wiring*: Relation with LSTMs

The incorporation of a Delay Line extends the memory of the model in a particular way, by forcing the persistence of an explicit representation of the input for a certain time. This technique relates to Long Short Term Memory (LSTM) networks Hochreiter and Schmidhuber [1997], a variant of RNNs that also augments recurrent networks with a 'memory cell'.

In LSTMs, the hidden layer is extended to incorporate a circuitry of *gates* that control the flow of information. The core of such circuitry is the *memory cell*, the component that maintains a persistent copy of the input over time. The gates add or remove information from the cell state through linear operations. Thus, the gates are designed to learn which information should be forgotten or retrieved, while the state cell acts like a conveyor belt on which the input is propagated over time.

Both the delay line and the gated memory in LSTMS are incorporated to the networks to ensure persistence of the input by copying it over timesteps. However, the techniques also exhibit notable differences in how the information is propagated: while in LSTMs, the information in the cell state is only modified with linear operations, the delay line incorporates a non-linear activation function (*tanh*) that has the effect of applying a form of decay over the input, such that the input values are closer to zero after every timestep.

These models also exhibit a different approach to forgetting. As mentioned, my model incorporates some form of forgetting by decaying the input over time. This decay is applied uniformly to all the units participating in the representation of a certain input vector. In contrast, in LSTMs the information does not uniformly and constantly deteriorate over time. Instead, the gate circuitry incorporates a *forget* gate that is trained to decide which information of the cell state should decay or be removed. The gate operates over units rather than over vectors, so the features of the input are not uniformly surpressed.

7.7.4 *Pre-training*: Relation with Dropout

The Incremental Novelty Exposure training regime of section 7.6, also finds a counterpart in the current deep learning literature, in the regularization technique of Dropout (Hinton et al. 2012, Srivastava et al. 2014). The idea behind Dropout is that, by randomly removing some of the units of the network during training, the model is forced to find solutions that do not rely too much on concrete correlations. This intuition is similar to the one that guides the INE: by constantly adding some novelty to the data, I force the network to not rely too much on accidental correlations. The crucial difference between both techniques is that Dropout modifies the architecture of the network to achieve this purpose, while the INE achieves a similar effect by manipulating the training data.

7.8 Conclusion

The Marcus et al. publication conveyed two main statements: first, that infants spontaneously extract identity rules from a short familiarization, and second, that neural networks are doomed to fail in such simple task. My work suggests that both the initial optimism for the generalization abilities of infants and the pessimism towards neural networks were overstated.

This study investigates the initial conditions that a neural network model should implement in order to account for the experimental results. First, I have proposed that networks should be *pre-wired* to incorporate a bias towards memorization of the input, which I have implemented as a delay line. With such *pre-wiring*, the model yields a notable proportion of significantly different responses between grammars. But despite such apparent success, the accuracy of the model in syllable prediction is very low.

Therefore, even though the successful discrimination between grammars is generally understood as abstraction of the underlying rule, my results show that significantly different responses can easily be found in a model that has not perfectly learnt such a rule. The corollary is that the generalization abilities of infants may have been overestimated; as a matter of fact, null results in similar experiments also point in that direction (see for instance footnote 1 in Gerken [2006]; also Geambasu&Levelt, p.c.).

But can neural networks go beyond grammar discrimination and accurately predict the next syllable according to a generalized rule? This work shows that this can be achieved when prior experience is incorporated in the model. I have hypothesized that, from all the information available in the environment, it is the gradual exposure to novel items that enhances generalization. This particular hypothesis deviates from related studies in which (i) an SRN was pre-trained to learn the relation of *sameness* between syllables [Seidenberg and Elman, 1999a], and (ii) an SRN was pre-trained with a set of sentences generated from a uniformly sampled vocabulary. Although the data used in the pre-training proposed here is less realistic than that used by Altmann [2002], my evaluation method is more strict (since I aim to test for accuracy rather than discrimination); for this reason, I first need to evaluate a model with a more constrained input. The next step in future work should be to explore whether the same results can be obtained when input data involving gradual novelty is generated from a grammar unrelated to the actual experiment.

Finally, from the perspective of the symbol vs. associations debate, at some abstract level of description, the delay line may be interpreted as providing the model with variables (that is, the dedicated group of nodes that encode the input at a certain time may be seen as a register) and the symbolic operation of "copy". It should be noted though that these groups of nodes are not isolated, and therefore, the learning algorithm needs to discover the structure in order to make use of it. Furthermore, it is uncontroversial that items are kept in memory for a certain lapse of time, so this structure is unlikely to constitute the core of the symbolic enterprise. If nevertheless my model is seen as compatible with the theory of rules-over-variables, my approach may be seen as providing a unifying model in which both sides in the debate can see their proposals reflected.

Conclusions

Chapter 8

Conclusions

8.1 Summary

The goal of this dissertation was to use the methodology offered by computational modelling to advance the current knowledge on the cognitive mechanisms behind ALL experiments. In my approach to the challenge, I conceptualized the process of learning the basic rules of a language as consisting of three steps: (i) memorization of sequence segments, (ii) computing the propensity to generalize, and (iii) generalization. In this dissertation I have proposed an account of each of these steps with a computational model.

Step (i) is relevant to understand how individuals segment a speech stream. In chapter 3 I have proposed R&R, a processing model that explains segmentation as a result of retention and recognition. This model offered an intuitive explanation of the process, and prompted the discovery that the memorization of segments tends to follow a skewed distribution rather than one that clearly separates statistically coherent items (words) from other segments of similar length but less statistically salient (partwords).

For step (ii), I propose that Simple Good Turing [Good, 1953], an existing *smoothing* model used in Natural Language Processing to account for unseen words in corpora, can be taken as a rational model. The principle it is based on —that the number of useful unseen items can be estimated from the number of times that items are seen once, twice, thrice, etc.— can explain the responses of individuals in the experiments.

As for step (iii), I first presented an extensive critical review of the existing models (chapter 6), in order to assess the state of the art and the critical issues that have not been resolved yet. After listing a desiderata for future research on generalization, I present a neural network model that emphasizes the role of the initial state of the model, based on two core ideas: *pre-wiring* the connections of the network to provide it with another type of memory, and *pre-training* to account for the relevant experience that influences generalization (concretely, experience that gradually incorporates novelty).

8.2 Contributions I: Exploring better ways to evaluate models

As reflected in § 2.4, there are no *a priori* criteria to determine what constitutes a good model. In this dissertation I have employed and developed different forms of evaluation.

To begin with, in chapter 5 I use *model sequencing* to study the behavior of models when combined in a pipeline sequence –that is, the output of one model (in this case the output of the R&R model) is used as the input of the next model (SGT). As reported, thanks to the contribution of both models the final output coincides with the empirical data.

In chapter 4 I used *model parallelisation* instead, to compare a range of models based on their goodness of fit to data from a number of experiments. As reported in that chapter, also for model parallelisation we can find different types of evaluation, roughly divided on those that assess the internal representations built by models while processing the stimuli, or those based instead on the 'behavioural responses' of the model. These are complementary forms of evaluation, but as shown, evaluation based on certain aspects of internal representations (in this case frequency distributions) can be necessary to differentiate between models.

Finally, in chapter 7 I evaluated my model based on two criteria: whether the model produced 'responses' that are sufficiently different between conditions, or whether it predicted the 'correct' outcome. The former was applied with the aim of reproducing the empirical results, while the latter aimed to assess the accuracy of the model, in order to address a more general theoretical question about whether perfect generalization in this model can be achieved.

After exploring these various forms of evaluation, what becomes clear is that each study may require a different approach, but in all cases modellers should strive to be very strict and creative in defining evaluation criteria that challenges models sufficiently to be able to distinguish between alternative models.

8.3 Contributions II: Exploring complementary levels of analysis

The models proposed throughout this dissertation are pitched at different levels of analysis. As advanced in chapter 2, each level suits different objectives.

To begin with, R&R is a processing level model. The main reason for this choice is that, in segmentation, the key open questions are about differences in behaviour between age groups and species. These differences cannot be investigated with rational models, since those models do not incorporate any properties of the cognitive system. Likewise, addressing these questions at the neural level is not likely to be helpful, since we require interpretable theories before addressing which implementational properties are responsible for the observed differences. And indeed, after proposing this model we have an intuitive theory of segmentation, based on basic perceptual and memory processes of retention and recognition.

In generalization, I identified a misconception in the field: namely, that statistical models should always generalize more when presented with more data. I have used a rational model (SGT, [Good, 1953]) to show that, when accounting for the propensity to generalize, this premise turns out to be false. This approach served as a first step to conceptualize the problem, and it paves the way for future models at the processing or implementational level.

Finally, I have proposed an implementational model of generalization (chapter 7). The goal of this model was precisely to understand how behaviour that appears symbolic can be accounted for at the implementational level, without the use of rules and symbols; hence this level of analysis was called for. Additionally, as seen in the corresponding chapter, the symbolic models proposed to explain the experiment in Marcus et al. [1999] are very simple: when assuming predefined rules over existing variables, the models do not offer much room for unexpected findings or predictions.

These three models target different types of questions, and it has been clearly useful to adhere to different levels of analysis depending on each particular goal.

8.4 Contributions III: Reframing key theoretical debates

One of the main advantages of the computational modelling methodology is that it forces researchers to be precise about all the details in a theory, and hence it helps clarifying misunderstandings. There are two important debates to which this dissertation has contributed.

First of all, with the proposal of a model for the propensity to generalize (chapter 5), I contributed to the debate about one vs. multiple mechanisms. Roughly speaking, most ALL studies can be categorized in one of these positions: those that postulate that the same mechanism for segmentation can be extended to account for generalization, and those which argue that we need to postulate an additional mechanism of a different nature. With my work, I have shown that one of the arguments for the latter position —namely, that a single statistical mechanism would predict better generalization with longer exposure— was a result of lack of formalization; a simple rational model like Simple Good Turing sufficed to illustrate this point.

Second, I have also contributed to a prominent debate about the allegedly symbolic nature of the mechanism of generalization. As reviewed in chapter 6, the debate condenses in disagreement on the cognitive realism of rules and variables (although I identified other questions to which computational models can contribute). The neural network model proposed in chapter 7 shows that a non-symbolic model can *learn* to use the information in a way that is descriptively equivalent to the use of rules over variables, even though at the level of implementation there are no rules or variables.

8.5 Future work

This dissertation has been structured along the three-step approach, which is a high-level conceptualization of the learning processes that take place in ALL experiments. Even though I have proposed a model for each step, there are details about the whole approach that are still underspecifed.

For instance, does the flow of information always proceed as a pipeline, or are steps (ii) and (iii) being constantly updated while step (i) is taking place? This may appear irrelevant if the information flows unidirectionaly, but we do not know if that is actually the case. For instance, the generalizations derived in step (iii) may have an influence on what is being memorized in step (i). These are open questions that this dissertation does not address and thus they should be tackled in future work.

Likewise, a more integrated model of the three steps is missing: while in chapter 5 I have used *model sequencing* to show that the output of step (i) has the necessary properties for step (ii) to reproduce the empirical data, this type of serialization is missing for connecting steps (ii) and (iii). Although the conceptualization makes it clear that the propensity to generalize is quantified as a probability for unseen items that should be divided between the generalized items computed in step (iii), how this probability is actually divided among these items has not been formalized yet.

Another direction that should be further explored is to extend the cross-species analysis with other animal species. In chapter 3 I presented simulations based on ALL experiments with rats, but as mentioned in the introduction (chapter 1), there exists a large body of ALL experiments involving other animals, such as birds and primates. These experiments are often not directly comparable with those on humans (e.g. experiments with songbirds generally require extensive training with feedback, and this contrasts especially with human infant experiments in which learning is typically spontaneous after short familiarization), but computational simulations may be used in creative ways to investigate empirical results in different species.

Finally, even in the case of humans, the datasets that have been studied in this dissertation are relatively small. This limitation is most serious in the case of generalization (step (iii)), for which I have focused exclusively on the data from Marcus et al. [1999] despite the existence of other experiments (see chapter 1 for other examples of generalization experiments in ALL). The reason for this is that the Marcus et al. experiment offers a very clean and illustrative dataset, and in spite of the replication issues, it is conceptually useful to discuss the challenges of generalization. Nonetheless, the predictions of models of generalization should be tested against other existing datasets to move forward on this question.

Bibliography

K. Abe and D. Watanabe. Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature neuroscience*, 14(8):1067–1074, 2011.

R. G. Alhama and W. Zuidema. Generalization in artificial language learning: Modelling the propensity to generalize. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 64–72, 2016. URL `http://anthology.aclweb.org/W16-1909`.

R. G. Alhama and W. Zuidema. Pre-wiring and pre-training: What does a neural network need to learn truly general identity rules? *Under review*, 2017a.

R. G. Alhama and W. Zuidema. Segmentation as retention and recognition: the R&R model. *Proceedings of the 39th Annual Conference of the Cognitive Science Society.*, 2017b.

R. G. Alhama and W. Zuidema. Computational models of rule learning. *To be submitted.*, 2017c.

R. G. Alhama, R. Scha, and W. Zuidema. *Rule Learning in Humans and Animals*, chapter 49, pages 371–372. 2014.

R. G. Alhama, R. Scha, and W. Zuidema. How should we evaluate models of segmentation in artificial language learning? In *Proceedings of 13th International Conference on Cognitive Modeling*, 2015.

R. G. Alhama, R. Scha, and W. Zuidema. Memorization of sequence-segments by humans and non-human animals: the Retention-Recognition model. *ILLC Prepublications*, PP-2016-08, 2016.

A. Alishahi and S. Stevenson. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5):789–834, 2008.

G. T. Altmann. Learning and development in neural networks–the importance of prior experience. *Cognition*, 85(2):B43–B50, 2002.

G. T. Altmann and Z. Dienes. Rule learning by seven-month-old infants and neural networks. *Science*, 284(5416):875–875, 1999.

J. R. Anderson. *Rules of the mind.* Psychology Press, 2014.

R. N. Aslin, J. R. Saffran, and E. L. Newport. Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4):321–324, 1998.

C. Bannard, E. Lieven, and M. Tomasello. Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41):17284, 2009.

J. Batali. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In *Linguistic evolution through language acquisition: Formal and computational models*. Citeseer, 1999.

G. J. Beckers, J. J. Bolhuis, K. Okanoya, and R. C. Berwick. Birdsong neurolinguistics: Songbird context-free grammar claim is premature. *Neuroreport*, 23(3):139–145, 2012.

R. Bod. Exemplar-based syntax: How to get productivity from examples. *The linguistic review*, 23(3):291–320, 2006.

G. Borensztajn, W. Zuidema, and R. Bod. Children's grammars grow more abstract with age—evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, 1(1):175–188, 2009.

J. S. Bowers. On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological review*, 116 (1):220, 2009.

J. S. Bowers. More on grandmother cells and the biological implausibility of PDP models of cognition: A reply to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010). *Psychological Review*, 2010.

C. E. Carr and M. Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Sciences*, 85(21):8311–8315, 1988.

N. Chater and P. Vitányi. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19 – 22, 2003. ISSN 1364-6613. doi: http://dx.doi.org/10.1016/S1364-6613(02)00005-0. URL `http://www.sciencedirect.com/science/article/pii/S1364661302000050`.

R. Chaudhuri and I. Fiete. Computational principles of memory. *Nature neuroscience*, 19(3):394–403, 2016.

N. Chomsky. *Syntactic Structures*. Mouton, The Hague, 1957.

N. Chomsky. *Aspects of the Theory of Syntax*, volume 119. MIT Press (MA), 1965.

N. Chomsky. Rules and representations. *Behavioral and Brain Sciences*, 3(01):1–15, 1980.

N. Chomsky. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group, 1986.

M. Christiansen, C. Conway, and S. Curtin. A connectionist single mechanism account of rule-like behavior in infancy. In *Proceedings of the 22nd annual conference of the cognitive science society*, pages 83–88, 2000.

M. H. Christiansen and S. Curtin. Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Sciences*, 3(8):289 – 290, 1999. ISSN 1364-6613. doi: http://dx.doi.org/10.1016/S1364-6613(99)01356-X. URL `http://www.sciencedirect.com/science/article/pii/S136466139901356X`.

M. H. Christiansen, J. Allen, and M. S. Seidenberg. Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, 13(2-3):221–268, 1998.

A. Clark and S. Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons, 2010.

E. V. Clark. *First language acquisition*. Cambridge University Press, 2009.

A. Cleeremans and J. L. McClelland. Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120(3):235, 1991.

F. Crick. The recent excitement about neural networks. *Nature*, 337:129–132, 1989.

W. Croft. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA, 2001.

W. Daelemans, S. Gillis, and G. Durieux. The acquisition of stress: A data-oriented approach. *Computational Linguistics*, 20(3):421–451, 1994.

Z. Dienes, G. Altmann, and S.-J. Gao. Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, 23(1):53–82, 1999.

P. F. Dominey. Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological cybernetics*, 73(3):265–274, 1995.

P. F. Dominey and F. Ramus. Neural network processing of natural language: I. sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, 15(1):87–127, 2000.

P. Eimas. Do infants learn grammar with algebra or statistics? *Science*, 284(5413): 435, 1999.

J. Elman. Generalization, rules, and neural networks: A simulation of marcus et. al. (1999). Html document., 1999.

J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

J. L. Elman. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press, 1998.

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

A. Endress and L. Bonatti. Words, rules, and mechanisms of language acquisition. *WIREs Cognitive Science*, 2016. (in press).

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

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

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

S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. *Advances in Neural Information Processing Systems 2*, pages 524–532, 1990.

C. J. Fillmore, P. Kay, and M. C. O'connor. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538, 1988.

W. T. Fitch and M. D. Hauser. Computational constraints on syntactic processing in a nonhuman primate. *Science*, 303(5656):377–380, 2004.

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71, 1988.

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

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

M. C. Frank, S. Goldwater, T. L. Griffiths, and J. B. Tenenbaum. Modeling human performance in statistical word segmentation. *Cognition*, 2010.

S. L. Frank and M. Čerňanskỳ. Generalization and systematicity in echo state networks. In . V. S. B.C. Love, K. McRae, editor, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 733–738. Cognitive Science Society, 2008.

R. M. French and G. W. Cottrell. Tracx 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

R. M. French, C. Addyman, and D. Mareschal. Tracx: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, 118(4):614, 2011.

R. L. Frost and P. Monaghan. Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, 147:70–74, 2016.

W. Gale and G. Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

L. Gamut. *Logic, Language, and Meaning: Intensional logic and logical grammar*. University of Chicago Press, 1991.

M. Gasser and E. Colunga. Babies, variables, and relational correlations. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society: August 13-15, 2000, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA*, volume 22, page 160. Lawrence Erlbaum Associates, 2000.

W. S. Geisler. Ideal observer analysis. *The visual neurosciences*, pages 825–837, 2003.

T. Gentner, K. Fenn, D. Margoliash, and H. Nusbaum. Recursive syntactic pattern learning by songbirds. *Nature*, 440(7088):1204–1207, 2006.

L. Gerken. Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, 98(3):B67 – B74, 2006. ISSN 0010-0277.

E. M. Gold. Language identification in the limit. *Information and control*, 10(5): 447–474, 1967.

A. Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

S. Goldwater, T. L. Griffiths, and M. Johnson. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the annual meeting of the association for computational linguistics*, volume 44, pages 673–680, 2006.

S. Goldwater, T. L. Griffiths, and M. Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54, 2009.

R. L. Gómez. Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436, 2002.

R. L. Gomez and L. Gerken. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70(2):109–135, 1999.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

T. L. Griffiths, N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in cognitive sciences*, 14(8):357–364, 2010.

R. Gómez and J. Maye. The developmental trajectory of nonadjacent dependency learning. *Infancy*, 7(2):183–206, 2005. ISSN 1532-7078.

M. D. Hauser, E. L. Newport, and R. N. Aslin. Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, 78(3): B53–B64, 2001.

D. O. Hebb. *The organization of behavior: A neuropsychological theory*. John Wiley & Sons, 1949.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

J. E. Hummel. Getting symbols out of a neural architecture. *Connection Science*, 23 (2):109–118, 2011.

R. Jackendoff. *Foundations of language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press, 2003.

H. Jaeger. The ‘‘echo state’’ approach to analysing and training recurrent neural networks. Technical report, German National Research Center for Information Technology, 2001.

H. Jaeger. Short term memory in echo state networks. Technical report, German National Research Center for Information Technology, 2002.

H. Jaeger. Echo state network. *Scholarpedia*, 2(9):2330, 2007.

L. A. Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948.

M. F. Joanisse and J. L. McClelland. Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(3):235–247, 2015.

E. Keuleers and W. Daelemans. Memory-based learning models of inflectional morphology: A methodological case-study. *Lingue e linguaggio*, 6(2):151–174, 2007.

S. Kirby, M. Dowman, and T. L. Griffiths. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245, 2007. doi: 10.1073/pnas.0608222104. URL `http://www.pnas.org/content/104/12/5241.abstract`.

D. C. Knill and A. Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences*, 27(12):712–719, 2004.

S. E. Kuehne, D. Gentner, and K. D. Forbus. Modeling infant learning via symbolic structural alignment. In *Proceedings of the twenty-second annual conference of the cognitive science society*, pages 286–291, 2000.

C. Kurumada, S. C. Meylan, and M. C. Frank. Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127(3):439–453, 2013.

M. D. Lee and E.-J. Wagenmakers. *Bayesian cognitive modeling: A practical course*. Cambridge University Press, 2014.

E. V. Lieven, J. M. Pine, and G. Baldwin. Lexically-based learning and early grammatical development. *Journal of child language*, 24(01):187–219, 1997.

R. D. Luce. Detection and recognition. In *Handbook of mathematical psychology*. New York: Wiley, 1963.

M. Lukoševičius. A practical guide to applying echo state networks. In *Neural Networks: Tricks of the Trade*, pages 659–686. Springer, 2012.

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.

B. MacWhinney. *The CHILDES project: Tools for analyzing talk*, volume 2. Psychology Press, 2000.

B. MacWhinney. A multiple process solution to the logical problem of language acquisition. *Journal of child language*, 31(04):883–914, 2004.

G. Marcus. *The algebraic mind.* Cambridge, MA: MIT Press, 2001.

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

G. F. Marcus. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3): 243–282, 1998.

G. F. Marcus. Connectionism: with or without rules?: Response to J.L. McClelland and D.C. Plaut (1999). *Trends in Cognitive Sciences*, 3(5):168 – 170, 1999a. ISSN 1364-6613. doi: http://dx.doi.org/10.1016/S1364-6613(99)01321-2. URL `http://www.sciencedirect.com/science/article/pii/S1364661399013212`.

G. F. Marcus. Do infants learn grammar with algebra or statistics? Response to Seidenberg and Elman, Negishi and Eimas. *Science*, 284:436–37, 1999b.

G. F. Marcus. Reply to Christiansen and Curtin. *Trends in Cognitive Sciences*, 3(8):290 – 291, 1999c. ISSN 1364-6613. doi: http://dx.doi.org/10.1016/S1364-6613(99)01358-3. URL `http://www.sciencedirect.com/science/article/pii/S1364661399013583`.

G. F. Marcus. Reply to Seidenberg and Elman. *Trends in Cognitive Sciences*, 3(8):288, 1999d.

G. F. Marcus. Rule Learning by seven-month-old infants and neural networks. Response to Altmann and Dienes. *Science*, 284:875, 1999e.

D. Mareschal and R. M. French. A connectionist account of interference effects in early infant memory and categorization. In *Proceedings of the 19th Annual Cognitive Science Society Conference*, pages 484–489, 1997.

D. Marr. *Vision. A computational investigation into the human representation and processing of visual information.* W. H. Freeman, New York, 1982.

J. L. McClelland. The place of modeling in cognitive science. *Topics in Cognitive Science*, 1(1):11–38, 2009.

J. L. McClelland and J. L. Elman. The TRACE model of speech perception. *Cognitive psychology*, 18(1):1–86, 1986.

J. L. McClelland and D. C. Plaut. Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3(5):166–168, 1999.

J. L. McClelland, M. M. Botvinick, D. C. Noelle, D. C. Plaut, T. T. Rogers, M. S. Seidenberg, and L. B. Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356, 2010.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

M. Negishi. Do infants learn grammar with algebra or statistics? *Science*, 284(5413): 435, 1999.

A. Newell. *A unified theory of cognition.* Harvard University Press, 1990.

E. L. Newport and R. N. Aslin. Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2):127–162, 2004.

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

L. Onnis, P. Monaghan, K. Richmond, and N. Chater. Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, 53(2):225–237, 2005.

B. Opitz and J. Hofmann. Concurrence of rule- and similarity-based mechanisms in artificial grammar learning. *Cognitive Psychology*, 77:77 – 99, 2015. ISSN 0010-0285. doi: http://dx.doi.org/10.1016/j.cogpsych.2015.02.003.

R. C. O'Reilly. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural computation*, 8(5):895–938, 1996.

R. C. O'Reilly. Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11):455 – 462, 1998. ISSN 1364-6613. doi: http://dx.doi.org/10.1016/S1364-6613(98)01241-8. URL `http://www.sciencedirect.com/science/article/pii/S1364661398012418`.

F. Pereira. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London*, 358(1769):1239–1253, 2000.

A. Perfors, J. B. Tenenbaum, T. L. Griffiths, and F. Xu. A tutorial introduction to bayesian models of cognitive development. *Cognition*, 120(3):302–321, 2011.

P. Perruchet and A. Vinter. Parser: A model for word segmentation. *Journal of Memory and Language*, 39(2):246–263, 1998.

P. Perruchet, M. D. Tyler, N. Galland, and R. Peerman. Learning nonadjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology*, 2004.

M. Peña, L. Bonatti, M. Nespor, and J. Mehler. Signal-driven computations in speech processing. *Science*, 298(5593):604–607, 2002.

S. Pinker. *The Language Instinct*. New York, NY: Harper Perennial Modern Classics, 1994.

S. Pinker. *Words and rules: The ingredients of language*. Basic Books, 2015.

S. Pinker and A. Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193, 1988.

D. C. Plaut and J. L. McClelland. Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review*, 117:284–290, 2010.

K. Plunkett and V. Marchman. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69, 1993.

Z. Pylyshyn. *Rules and representations: Chomsky and representational realism*, pages 231–51. Blackwell Pub, 1991.

R. Quian Quiroga and G. Kreiman. Measuring sparseness in the brain: comment on Bowers (2009). 2010.

A. S. Reber. Implicit learning of artificial grammars. *"Journal of Verbal Learning and Verbal Behavior*, 5:855–863, 1967.

T. Regier and S. Gahl. Learning the unlearnable: the role of missing evidence. *Cognition*, 93(2):147 – 155, 2004. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2003.12.003. URL `http://www.sciencedirect.com/science/article/pii/S0010027704000587`.

D. L. Rohde and D. C. Plaut. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109, 1999.

D. Rumelhart and J. McClelland. On learning past tenses of English verbs. In D. Rumelhart and J. McClelland, editors, *Parallel Distributed Processing, Vol. 2*, pages 318–362. MIT Press, Cambridge, MA, 1986.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996a.

J. R. Saffran, E. L. Newport, and R. N. Aslin. Word segmentation: the role of distributional cues. *Journal of Memory and Language*, 35(4):606–621, 1996b.

J. R. Saffran, E. L. Newport, R. N. Aslin, R. A. Tunick, and S. Barrueco. Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8(2):101–105, 1997.

A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016.

R. Scha. *Computertoepassingen in de Neerlandistiek.*, chapter Taaltheorie en taaltechnologie; competence en performance., pages 7–22. LVVM, Almere., Almere, 1990.

L. A. Segel and L. Edelstein-Keshet. *A Primer in Mathematical Models in Biology*, volume 129. SIAM, 2013.

M. S. Seidenberg and J. L. Elman. Do infants learn grammar with algebra or statistics? *Science*, 284(5413):433, 1999a.

M. S. Seidenberg and J. L. Elman. Networks are not ‘hidden rules’. *Trends in Cognitive Sciences*, 3(8):288–289, 1999b.

M. S. Seidenberg and J. L. McClelland. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523, 1989.

D. Servan-Schreiber, A. Cleeremans, and J. L. McClelland. Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, 7(2-3):161–193, 1991.

L. Shastri and S. Chang. A spatiotemporal connectionist model of algebraic rule-learning. Technical Report TR-99-011, Berkeley, California: International Computer Science Institute, 1999.

L. Shastri, V. Ajjanagadde, L. Bonatti, T. Lange, and M. Dyer. From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 19(2): 326–337, 1993.

T. R. Shultz. Rule learning by habituation can be simulated in neural networks. In *Proceedings of the twenty first annual conference of the Cognitive Science Society*, pages 665–670, 1999.

T. R. Shultz. Assessing generalization in connectionist and rule-based models under the learning constraint. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*, 2001.

K. Sima'an. *Learning Efficient Disambiguation*. PhD thesis, ILLC dissertation series 1999-02, 1999.

S. Sirois, D. Buckingham, and T. R. Shultz. Artificial grammer learning by infants: an auto-associator perspective. *Developmental Science*, 3(4):442–456, 2000.

R. Skousen. Analogy: A non-rule alternative to neural networks. *Rivista di linguistica*, 7:213–232, 1995.

R. Skousen, D. Lonsdale, and D. B. Parkinson. *Analogical modeling: An exemplar-based approach to language*, volume 10. John Benjamins Publishing, 2002.

K. Smith. Iterated learning in populations of bayesian agents. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 697–702, 2009.

M. Spierings, A. de Weger, and C. Ten Cate. Pauses enhance chunk recognition in song element strings by zebra finches. *Animal cognition*, 18(4):867–874, 2015.

M. J. Spierings and C. ten Cate. 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, 2016.

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

L. Steels. Fluid construction grammar. In *The Oxford Handbook of Construction Grammar*. 2013.

D. G. Stork. Is backpropagation biologically plausible? In *International 1989 Joint Conference on Neural Networks*, pages 241–246 vol.2, 1989. doi: 10.1109/IJCNN. 1989.118705.

D. Swingley. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50(1):86–132, 2005.

J. B. Tenenbaum and T. L. Griffiths. Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, 24(04):629–640, 2001.

E. D. Thiessen and J. R. Saffran. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, 39(4): 706, 2003.

E. D. Thiessen and J. R. Saffran. Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language learning and development*, 3(1): 73–100, 2007.

B. Thompson, S. Kirby, and K. Smith. Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, 113(16):4530–4535, 2016.

M. Tomasello. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1-2):61–82, 2001.

J. M. Toro and J. B. Trobalón. Statistical computations over a speech stream in a rodent. *Perception & psychophysics*, 67(5):867–875, 2005.

S. Tsuji, C. Bergmann, and A. Cristia. Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, 9(6):661–665, 2014.

C. A. Van Heijningen, J. De Visser, W. Zuidema, and C. Ten Cate. Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences*, 106(48):20538–20543, 2009.

M. Vilcu and R. F. Hadley. Two apparent 'counterexamples' to Marcus: A closer look. *Minds and Machines*, 15(3):359–382, 2005.

X. Xie and H. S. Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 15(2):441–454, 2003.

D. Zipser and R. A. Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331 (6158):679–684, 1988.

W. Zuidema. What are the productive units of natural language grammar?: a DOP approach to the automatic identification of constructions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 29–36. Association for Computational Linguistics, 2006.

W. Zuidema and B. de Boer. Modeling in the language sciences. *Research Methods in Linguistics*, page 422, 2014.

W. H. Zuidema. How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 51–58. MIT Press, 2003.

Samenvatting

In de afgelopen twintig jaar is *Artificial Language Learning* één van de belangrijkste benaderingen geworden in het onderzoek naar hoe we het spraaksignaal in stukken opdelen en hoe we regels van onze moedertaal leren. In de experimenten in dit paradigma krijgen proefpersonen een reeks van stimuli te horen of te zien, waarin bepaalde (statistische) regelmatigheden verborgen zitten die lijken op de regelmatigheden in spraak en taal. Proefpersonen worden vervolgens getest om er achter te komen òf, en onder welke voorwaarden zij het onderliggende patroon kunnen ontdekken.

In dit proefschrift gebruik ik computermodellen om de resultaten van dergelijke experimenten — met babies, volwassenenen en dieren — beter te analyseren en interpreteren. Doel van die analyse is een beter begrip van de meest fundamentele mechanismen die een rol spelen bij het leren van taal. Ik stel voor om over het leerproces in *Artificial Language Learning*-experimenten na te denken als bestaande uit 3 stappen: (i) het opslaan in het geheugen van fragmenten van de input; (ii) het berekenen van de bereidheid tot generalisatie; (iii) het daadwerkelijk generaliseren. In dit proefschrift werk ik voor ieder van deze stappen een computermodel uit.

Stap (i) is relevant voor ons begrip van hoe mensen het continue spraaksignaal in discrete segmenten kunnen opdelen. In hoofdstuk 3 stel ik het *Rentention & Recognition*-model voor, dat het segmentatie-proces beschrijft als het resultaat van een interactie tussen kortdurende *retentie* van stukken van het signaal, opslag in het geheugen en het *herkennen* van een segment in de input. Ik laat zien dat dit model een verklaring biedt voor een reeks van empirische resultaten met mensen en ratten (Peña et al., 2002; Toro and Trobalón, 2005; Frank et al., 2010). R&R geeft een natuurlijke verklaring voor het segmentatie-proces. Bovendien was R&R de aanleiding om eens goed te kijken naar de frequentie-verdelingen van segmenten, en daarmee tot de ontdekking dat die verdelingen vaak heel scheef zijn en dat de frequenties van zogeheten ‘words’ en ‘partwords’ vaak overlappen.

Stap (ii) is in dit proefschrift een aparte stap in het generalisatie-proces, en dat feit op zich is één van de innovaties in dit proefschrift. In hoofdstuk 5 gebruik ik een bestaand model uit de natuurlijke taalverwerking voor het ‘gladstrijken’ (‘smoothing’) van kansverdelingen, te weten het Simple Good Turing-model (SGT, Good (1953)). Ik laat zien dat het principe waar dit model op gebaseerd is een goede verklaring biedt voor de observaties uit experimenten met mensen.

In hoofdstuk 6 begin ik mijn analyse van stap (iii) met een kritisch overzicht van bestaande modellen, om een duidelijk beeld te krijgen van wat er op dit punt al bereikt is en wat de open vragen zijn. Ik eindig dat hoofdstuk met een lijst van 'desiderata' — nastrevenwaardige kenmerken van toekomstige modellen van generalisatie. In hoofdstuk 7 werk ik vervolgens een neuraal netwerk-model uit dat al voldoet aan een belangrijk deel van de genoemde kenmerken op dat verlanglijstje. Mijn model biedt een verklaring voor de resultaten van één van de meest invloedrijke studies uit de ALL-literatuur (Marcus et al., 1999). Het model maakt gebruik van twee cruciale ideeën: '*pre-wiring*' — het idee dat het netwerk al voorafgaand aan het leren verbindingen tussen neuronen heeft die het een extra vorm van geheugen geven — en '*pre-training*' — het idee dat het netwerk voordat het de stimuli uit het experiment te zien krijgt, al op een manier getraind is die generalisatie faciliteert.

Op meerdere plekken in het proefschrift komen methodologische kwesties aan de orde. Ik begin met een bespreking van Marr's analyse-niveaus (Marr, 1982) en de consequenties van de keuze voor een top-down of bottom-up benadering. Daarna bespreek ik de voor- en nadelen van die verschillende keuzes. Ik concludeer dat wat de beste benadering is, sterk afhangt van de onderzoeksvraag. Tenslotte behandel ik ook methodologische vragen over hoe we modellen moeten evalueren. In een vergelijkende studie van computermodellen in hoofdstuk 4 behandel ik een aantal alternatieve manieren van evalueren ('model parallelisation' vs. 'model sequencing', 'internal representations' vs. 'external output'). Daarbij laat ik zien dat verschillende evaluatieprocedures naast elkaar kunnen en moeten bestaan, en dat we voor de ALL-modellen striktere evaluatie-criteria kunnen formuleren.

Dit proefschrift biedt al met al een geïntegreerd perspectief op de segmentatie- en generalisatie-processen in *Artificial Language Learning*, op basis van computermodellen die empirische observaties reproduceren, toetsbare voorspellingen doen, een bijdrage leveren aan het oplossen van open vragen in het vakgebied, en daarmee tot een beter begrip leiden van de fundamentele processen die aan het leren van taal ten grondslag liggen.

Abstract

Artificial Language Learning has, over the last 20 years, become a key paradigm to study the nature of learning biases in speech segmentation and rule generalization. In experiments in this paradigm, participants are exposed to a sequence of stimuli that have certain statistical properties, and which may follow a specific pattern. The design intends to mimic particular aspects of speech and language, and participants are tested on whether and under which conditions they can segment the input and/or discover the underlying pattern.

In this dissertation, I have used computational modelling to interpret results from Artificial Language Learning experiments on infants, adults and even non-human animals, with the goal of understanding the most basic mechanisms of language learning. I have conceptualized the process as consisting of three steps: (i) memorization of sequence segments, (ii) computing the propensity to generalize, and (iii) generalization. Along the dissertation I have proposed an account of each of these steps with a computational model.

Step (i) is relevant to understand how individuals segment a speech stream. In chapter 3 I have proposed R&R, a processing model that explains segmentation as a result of retention and recognition. I have shown that this model can account for a range of empirical results on humans and rats (Peña et al., 2002; Toro and Trobalón, 2005; Frank et al., 2010). R&R offers an intuitive explanation of the segmentation process, and it also prompted the discovery that the memorization of segments tends to produce skewed and overlapping distributions of words and partwords.

Identifying step (ii) as a separate step is actually a contribution from this dissertation (as is explained in chapter 5). I propose that Simple Good Turing (or SGT, Good (1953)), an existing *smoothing* model used in Natural Language Processing to account for unseen words in corpora, can be taken as a rational model for step (ii) since the principle it is based on can explain the responses of individuals in the experiments.

As for step (iii), I first presented an extensive critical review of the existing models (chapter 6), in order to identify the state of the art and the critical issues that still have not been resolved. After listing desiderata for future research on generalization, I present a neural network model that addresses some of those. Concretely, my neural network model accounts for the results of one influential experiment (Marcus et al., 1999) by incorporating two core ideas: *pre-wiring* the connections of the network to provide the model with another type of memory, and *pre-training* to account for the relevant experience that influences generalization (concretely, incremental presentation of novelty).

Throughout the dissertation, I also reflect on methodological issues in computational modelling. After introducing Marr's levels of analysis (Marr, 1982) and discussing the implications of top-down and bottom-up approaches, I explore the strengths and weaknesses of each level of analysis with each one of the proposed models, and conclude that the best choice depends on the particular research question. Finally, I also discuss issues with model evaluation. Through a model comparison study (chapter 4), I explore alternative evaluation procedures (model parallelisation vs. model sequencing, internal representations vs. external output). This study illustrates the need for complementary types of analysis of empirical results, and it provides the basis for devising stricter evaluation criteria.

This dissertation thus provides an integrated account of segmentation and generalization, based on computational models that have been shown to reproduce empirical results, produce testable predictions, contribute to unresolved theoretical questions and, overall, increase our understanding of the basic processes of language learning.

Curriculum Vitae

Education

2012-2017 PhD in Computational Linguistics

Institute for Logic, Language and Computation (ILLC),
University of Amsterdam
Supervisors: Willem Zuidema, Remko Scha† and Carel ten Cate
Thesis: *Computational Modelling of Artificial Language Learning: Retention, Recognition and Recurrence.*

2010-2012 MSc in Cognitive Sciences and Language

Universitat de Barcelona
Supervisors: Antònia Martí and Xavier Carreras
Thesis: *Characterization of light verb constructions in Distributional Semantic Models*.

2003-2008 BSc+MSc in Computer Engineering

Universitat Autònoma de Barcelona
5-year degree (336 ECTS).

Publications

Alhama, R. G., & Zuidema, W. (2017) Segmentation as Retention and Recognition: the R&R Model. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, (CogSci2017).

Stanojević, M. & **Alhama, R. G.** (2017) Neural Discontinuous Constituency Parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (EMNLP2017).

Alhama, R. G., & Zuidema, W. (2016) Pre-Wiring and Pre-Training: What does a neural network need to learn truly general identity rules? In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches* (NIPS2016). Extended version under review for *Journal of Artificial Intelligence Research.*

Alhama, R. G., & Zuidema, W. (2016) Generalization in Artificial Language Learning: Modelling the Propensity to Generalize. In *Cognitive Aspects of Computational Language Learning* (CogACLL at ACL2016).

Alhama, R. G., Scha, R., & Zuidema, W. (2016) Memorization of sequence-segments by humans and non-human animals. *ILLC Prepublications, ILLC (University of Amsterdam), PP-2016-08.*

Alhama, R. G., Scha, R., & Zuidema, W. (2015) How should we evzaaluate models of segmentation in Artificial Grammar Learning? In *Proceedings of 13th International Conference on Cognitive Modeling*, (ICCM15). **Best poster award.**

Alhama, R. G., Scha, R., & Zuidema, W. (2014) Rule learning in humans and animals. In *Proceedings of the International Conference on the Evolution of Language* (EvolangX).

Martí, M. A., **Alhama, R. G.** & Recasens, M. (2012) Los avances tecnológicos y la ciencia del lenguaje. [Technological advances and the language sciences.] In T. Jiménez Juliá, B. López Meirama, V. Vázquez Rozas, and Alexandre Veiga (eds.), *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*, pages 543-553.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder... ": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective