This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Transmission, Induction, and Evolution

Bill Thompson

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy
University of Edinburgh
2014
Abstract

Many human behaviours are thought to depend upon cognitive capacities enriched with innate domain-specific knowledge. Underpinning this view is the hypothesis that evolution can shape cognition to include strong innate inductive biases. In this thesis, I re-examine that hypothesis with respect to a broad class of behaviours: those that we learn from other individuals. Taking human language as a test case, I present an analysis of the co-evolutionary process that underpins the formation of innate constraints on cognition for behaviours that are culturally transmitted through inductive inference. I derive a series of mathematical models of this process, built around Bayesian models of cognition and cultural transmission, and ask how they can inform our expectations about cognition in a cultural species. I argue that the traditional marriage of nativism and evolutionary reasoning is undermined by this process, as is the suggestion that cognitive adaptation to cultural behaviours is outright implausible. I explore the co-evolutionary dynamics induced by cultural transmission, and conclude that they can radically manipulate the evolution of cognition: culture can intervene in the formation of hard-wired knowledge, but nevertheless facilitate rapid cognitive adaptation. The analyses I report make strong, testable predictions about the nature of inductive biases for cultural behaviours, and offer solutions to a number of long-standing conundrums in the evolution of language.
Declaration

I declare that this thesis was composed by myself. The work contained therein is my own, except where explicitly stated otherwise in the text, and has not been submitted in any previous application for a degree.

(Bill Thompson)
Acknowledgements

It has been an enormous privilege to learn from Simon Kirby and Kenny Smith, who have offered patience, expertise, and friendship in equal measure: as supervisors they have transformed the way I think about language; as pals, many more important topics. Theirs is a wave of progress shared generously and without expectation: undertaking this research under their guidance has been hugely gratifying and, above all, good fun.

I am extremely grateful to my examiners, Rob Truswell and Tom Griffiths, for generously taking the time to read this thesis in detail, and for their invaluable insights and suggestions.

I have been fortunate to enjoy the help and guidance of numerous individuals during the preparation of this research: for kindly offering feedback on the co-evolutionary models of chapters three and four, I am hugely grateful to Nick Barton, Bart de Boer, Kevin Laland, Mark Pagel, Andrea Ravignani, Amy Perfors, Geoff Pullum, and Pete Richerson; for a crash course in set theory, and for maths help and advice in general, I am deeply indebted to Simon Haines and Andrea Ravingani; for teaching me the log-sum trick, I thank Carmen Saldaña; for all those books I could never have bought, but have read thanks to his generosity, I thank Jim Hurford; for several years of invaluable discussion, I thank Vanessa Ferdinand, Sean Roberts, Matt Spike, James Thomas, and the many students of Simulating Language, whose questions and insights have on many occasions forced me to sharpen my thinking about cultural transmission; for allowing me the freedom to pursue a topic that sits uncomfortably between discipline boundaries, I am grateful to the department of Linguistics and English Language at The University of Edinburgh; for innumerable kinds of help and advice, and for running the most efficient department in the world, I am indebted forever to Katie Keltie, Toni Noble, Lynsey Buchanan, and Eddie Douborg; for financial support, I thank the Engineering and Physical Sciences Research Council. This thesis would not exist were
it not for the company of the bafflingly smart, thoughtful, and creative individuals of the LEC who, despite these qualities, have been a wonderful bunch of friends and colleagues.

As ever, I thank my parents, Andrew and Judy Thompson, who make more of a difference every day than I do in a year, but have nevertheless shown me incomparable support and pride every step of the way. Living vicariously through the extraordinary experiences of my brother, Robbie Thompson, has kept me upbeat after many a long day during the writing of this thesis. Finally, for being my raison d’être, and everything else, I thank Laura, who has suffered my distraction with grace, and held my hand through every page.
# Contents

Abstract ii

Acknowledgements v

Contents viii

List of Tables ix

List of Figures x

Introduction 3

1 Nativism, Evolution, and Culture 4
   1.1 Nativism: an enduring thesis ........................................ 5
      1.1.1 Platonic Nativism .................................................. 5
      1.1.2 Continental Rationalism ........................................... 7
      1.1.3 Empiricism & the Enlightenment ................................. 8
   1.2 Contemporary Nativism .................................................. 11
      1.2.1 Linguistic Nativism ................................................ 13
   1.3 Evolution & Nativism ................................................... 25
      1.3.1 The Evolutionary Defense of Linguistic Nativism ............... 26
      1.3.2 Cultural Evolution ................................................ 33
   1.4 Summary ................................................................. 41

2 Co-evolution: A Review of Existing Models 43
   2.1 Gene-Culture Co-evolution ......................................... 44
      2.1.1 Social Learning .................................................... 45
2.1.2 Learned Birdsong
2.1.3 Boyd and Richerson’s Direct Bias
2.1.4 Deafness & Sign-languages
2.1.5 General Discussion of Gene-Culture Co-evolutionary Models
2.2 Co-evolutionary Models of Language
2.2.1 The Nowak Series
2.2.2 The Iterated Learning Series
2.2.3 The C&C Series
2.3 Summary
2.3.1 Model Desiderata
3 The Evolution of Inductive Biases for Culture: A Formal Analysis
3.1 A Co-evolutionary Model
3.1.1 A Model of Language Learning
3.1.2 A Model of Biological Inheritance
3.1.3 Cultural Transmission
3.1.4 Evolutionary Analysis
3.2 Results
3.2.1 Numerical Analysis & Initial Conditions
3.2.2 Two Baseline Analyses
3.2.3 Full Co-evolutionary Analyses
3.2.4 Stochastic Simulations in Finite Populations
3.3 Discussion
3.3.1 Evolution, Nativism, & Culture without Amplification
3.3.2 Evolution, Nativism, & Culture with Amplification
3.3.3 Relation to Existing Claims
4 Questioning the Basics: Alternative Co-evolutionary Analyses
4.1 Biological Evolution
4.1.1 Non-Arbitrary Languages (Asymmetric Fitness)
4.1.2 Fitness from Learning not Communication
4.1.3 Continuous Biological Types
4.2 Individual Learning
4.2.1 Interpolation between MAP & Sample

vii
List of Tables

1.1 A simple schema to distinguish various approaches to the cognitive architecture supporting language, adapted from Ullman et al. (2013). . . . . . 23

3.1 Comparison of Evolved Biases in the Sample Learner Model. . . . . . . 125
3.2 Comparison of Evolved Biases in the MAP model. . . . . . . . . . . . . . . 132
List of Figures

3.1 Mutation dynamics among genome classes (M). Black circular markers indicate the $ij^{th}$ coordinate. The colour of the surrounding square gives $m_{ij}$. 103

3.2 Evolution in the Biology Only model. Each row shows initial (red) and stable (blue) distribution of genome classes in the left hand column, and the change over time of the mean (and one standard deviation) genetic make-up of the population in the right hand column. In all cases extreme genomes come to dominate the population and remain stable thereafter. $n = 100, \mu = 0.001$. 114

3.3 Evolution in the No Learning No Culture model. Each row shows initial (red) and stable (blue) distribution of genome classes in the left hand column, and the change over time of the mean (and one standard deviation) genetic make-up of the population in the right hand column. In all cases extreme genomes come to dominate the population and remain stable thereafter. $n = 100, \mu = 0.001$. 117

3.4 Evolution in the sample learner model. Left hand column panels: initial genome class distributions ($g^0$) are not shown, since the final distributions ($g^*$) would be obscured. As before, the top row corresponds to Binomial initial conditions, the middle Neutral, and the bottom Nativised. Right hand column panels: the green line shows the mean prior bias ($\Phi$) in the population; one standard deviation in the distribution of biases in the population is shown above and below the green line by the shaded area; the purple line shows $x_1$, the population frequency of language $L_1$. $n = 100, \mu = 0.001, N = 2, p = .95$. 119

3.5 Evolution in the sample learner model. $n = 100, \mu = 0.001, N = 4, p = .95$. 123
3.6 Evolution in the sample learner model. \( n = 100, \mu = 0.001, N = 8, p = .95 \) 124
3.7 Evolution in the Sample learner model. \( n = 50, \mu = 0.001, N = 2, p = .95 \) 127
3.8 Evolution in the MAP learner model. \( n = 100, \mu = 0.001, N = 2, p = .95 \) 128
3.9 Evolution in the MAP learner model. \( n = 50, \mu = 0.001, N = 2, p = .95 \) 131
3.10 Evolutionary outcomes of stochastic simulations of the mathematical model in finite populations of learners. Markers show the average cognitive bias \( \Phi \) and the proportion \( x_1 \) of language \( L_1 \) speakers at generation 1000 for sample (yellow markers) and MAP (purple markers) populations. \( N = 100, n = 100, N = 2, \mu = .001 \) 133
3.11 Evolutionary outcomes of stochastic simulations of the mathematical model in finite populations of learners. Markers show the average cognitive bias \( \Phi \) and the proportion \( x_1 \) of language \( L_1 \) speakers at generation 5000 for sample (yellow markers) and MAP (purple markers) populations. \( N = 100, n = 50, N = 2, \mu = .001 \) 134
4.1 Evolution in MAP learner populations who receive a 1000-fold fitness advantage \( (\pi = 1000) \) for sharing language \( L_1 \). **Left hand column panels**: distributions of genome classes. As before, the top row corresponds to Binomial initial conditions, the middle Neutral, and the bottom Nativised. **Right hand column panels**: the green line shows the mean prior bias \( \Phi \) in the population; one standard deviation in the distribution of biases in the population is shown above and below the green line by the shaded area; the purple line shows \( x_1 \), the population frequency of language. \( n = 100, \mu = 0.001, N = 2, p = 0.95 \) 148
4.2 Evolution in sample learner populations who receive a fitness advantage for acquiring language \( L_1 \). \( n = 100, \mu = 0.001, N = 2, p = 0.95 \) 150
4.3 Mean bias \( \Phi^* \) and proportion of language \( L_1 \) users \( x_1^* \) in the equilibrium distributions for populations of Sample learners who receive a fitness advantage for acquiring language \( L_1 \) proportional to \( \pi \), over a range of values for \( \pi \). \( n = 100, \mu = 0.001, N = 2, p = 0.95 \), Binomial initial conditions. 151
4.4 Mean bias $\Phi^*$ and proportion of language $L_1$ users $x_1^*$ after 10000 generations in populations of 100 sample (yellow markers, 50 simulations) and MAP (purple markers, 50 simulations) learners. $n = 100, \mu = 0.001, N = 2, p = 0.95, N = 100$, Binomial initial conditions 153

4.5 Evolution in populations of sample learners whose fitness is assessed by their ability to correctly learn their teachers’ language. $n = 50, \mu = 0.001, N = 2, p = 0.95$ 156

4.6 Cognitive and linguistic outcomes after 5000 generations of co-evolution in populations of $N = 100$ sample learners (yellow markers) and MAP learners (purple markers) whose genomes are drawn from a continuous range of possibilities. $\beta = 50, \sigma = 0.01, \mu = 0.001, N = 2, p = 0.95$ 158

4.7 Example beta distributions 159

4.8 Cognitive and linguistic outcomes after 5000 generations of co-evolution in populations of $N = 100$ sample learners (yellow markers) and MAP learners (purple markers) whose genomes are drawn from a continuous range of possibilities, and face no cost to specialisation. $\beta = 1, \sigma = 0.01, \mu = 0.001, N = 2, p = 0.95$ 160

4.9 Evolutionary outcomes after 5000 generations under weak costs to specialisation in sample learner (yellow points) and MAP learner populations, for two of the beta distributions shown in figure 4.7. Panel (a) shows the outcome of simulations with very weak ($\beta = 2$) costs, while panel (b) shows those outcomes under stronger, but still relatively weak, costs ($\beta = 10$). $N = 100, u = 0.001, \sigma = 0.01, N = 2, p = 0.95$ 161

4.10 Mean bias $\Phi^*$ and proportion of language $L_1$ users $x_1^*$ in the equilibrium distributions for populations of learners whose induction probabilities follow posterior probabilities exponentiated by $r$, over a range of values for $r$. $n = 100, \mu = 0.001, N = 2, p = 0.95$, Binomial initial conditions 162

4.11 Evolution in populations of learners whose induction probabilities follow posterior probabilities exponentiated by $r = 2$. The green line and fill show the mean cognitive bias $\Phi$ and half of one standard deviation in that distribution. The purple line shows the proportion of the population speaking language $L_1$. $n = 100, \mu = 0.001, N = 2, p = 0.95$, Binomial initial conditions 163
4.12 Evolution in populations of learners whose induction probabilities follow posterior probabilities exponentiated by $r = 5$. The green line and fill show the mean cognitive bias $\Phi$ and half of one standard deviation in that distribution. The purple line shows the proportion of the population speaking language $L_1$. $n = 100$, $\mu = 0.001$, $N = 2$, $p = 0.95$, Binomial initial conditions.

4.13 Evolution in populations that contain both MAP and sample learners. In all panels: the purple line shows the proportion of sample learners in the population; the red line shows the mean innate bias $\Phi$ in the populations, and; the blue line shows $x_1$, the population frequency of speakers of language $L_1$. Plots show differing initial conditions with respect to the initial frequency of sample learners (All-None-Half) over the rows, and the mean innate bias of the first population over the columns (Neutral-Nativised $L_1$ -Nativised $L_0$). All conditions lead to the same outcome: MAP learners with weak innate biases supporting a linguistic universal.

4.14 (a): The distribution (probability density) of priors $\alpha$ under the assumption that genomes follow an exponential distribution parametrised with rate $\lambda = 5$: the distribution strongly favours the weakest biases. (b): The prior distribution (probability density) over $p$ for a range of values of $\alpha$: higher $\alpha$ implies a stronger inductive bias in favour of the ordering rule $S \rightarrow X Y$.

4.15 The distribution of final-population mean cognitive biases $\Phi$ over 100 simulations of 5000 generations in populations of MAP learners. Each bar gives the proportion of simulations in which the average final population bias fell within the range of the bin. The black dashed line shows the expected distribution of biases if there were no selection (scaled for easy comparison). $N = 100$, $\lambda = 5$, $\sigma = .1$, $\mu = 0.001$, $N = 4$.
4.16 The distribution of final-population mean cognitive biases $\Phi$ over 100 simulations of 5000 generations in populations of Sample learners. Each bar gives the proportion of simulations in which the average final population bias fell within the range of the bin. The black dashed line shows the expected distribution of biases if there were no selection (scaled for easy comparison). $N = 100, \lambda = 5., \sigma = .1, \mu = 0.001, N = 4$ 171

4.17 The distribution of final-population $p^*$ values over 100 simulations of 5000 generations in populations of Sample learners. Each bar shows the proportion of simulations in which the average of the final population’s $p$ values (which dictate the relative use of the two possible ordering patterns) fell within the range of the bin. $N = 100, \lambda = 5., \sigma = .1, \mu = 0.001, N = 4$ 172

4.18 Co-evolution in populations of sample learners that learn from multiple individuals. $n = 100, \mu = 0.001, N = 2, p = 0.95$ 175

4.19 Co-evolution in populations of sample learners that learn from multiple individuals. $n = 100, \mu = 0.001, N = 6, p = 0.95$ 177

4.20 Co-evolution in populations of MAP learners who learn from multiple individuals. $n = 100, \mu = 0.001, N = 2, p = 0.95$ 178

4.21 Co-evolution in populations of MAP learners who learn from multiple individuals. $n = 100, \mu = 0.001, N = 6, p = 0.95$ 179

5.1 Maximum likelihood values of $\hat{\alpha}$ for S&W’s data by majority marker input proportion. 200

5.2 Histogram of maximum likelihood values of $\hat{\alpha}$ for S&W’s data over 10000 samples of the data where $|\frac{1}{2}| = 12$ for all input marker proportion regimes $i$. 201

5.3 The logarithms of the mean values for $\hat{\alpha}$ by input regime over 10000 simulated frequency learning experiments in which simulated learners have a strong bias towards regularity. Errorbars give the logarithms of 95% confidence intervals in the raw $\hat{\alpha}$ samples. The dashed horizontal line illustrates the true $\alpha$. The solid horizontal line at 0 indicates the root of each bar, and the unbiased prior. $n^t = n^{t-1} = 12; 30$ participants per input regime per experiment, $\alpha = 0.1 (\ln(\alpha = 0.1) = -2.3))$ 203

xv
5.4 The logarithms of the mean values for $\bar{\alpha}$ by input regime over 10000 simulated frequency learning experiments in which simulated learners have a weak bias towards variability. Error-bars give the logarithms of 95% confidence intervals in the raw $\bar{\alpha}$ samples. The dashed horizontal line illustrates the true $\alpha$. The solid horizontal line at 0 indicates the root of each bar, and the unbiased prior. $n^t = n^{t-1} = 12$; 30 participants per input regime per experiment, $\alpha = 5$. ($\ln(\alpha) = 0.693$). 205

5.5 Honesty of various input regimes for $\alpha = 0.1$ (red) and $\alpha = 2$ (blue). 207

5.6 The Honesty of all possible input proportions averaged over the range of regularisation biases ($\bar{\alpha} = [0.001, 0.99]$). 209

5.7 A schematic presenting a range of methods for computing transition matrices from experimental data. A: Populating a transition matrix with empirically observed (input, output) datapoints. B: Computing a transition matrix from a model whose parameters have been fit to a complete dataset. C: Populating a transition matrix by fitting model parameters to individual participants, computing transition matrices for each, and averaging over/mixing those individual matrices to obtain the transition matrix corresponding to a well-mixed population containing these biases. D: Computing a transition matrix by fitting model parameters to individual participant data, inferring the distribution(s) of those parameters/biases in the participant population, integrating over the range of possible biases, mixing the transition matrices defined by each, while weighting their contributions according the inferred distribution. 216
Introduction

At the heart of contemporary cognitive science is the hypothesis that many human behaviours depend upon specialised cognitive capacities. Underpinning that hypothesis is the premise that the existence of hard-wired knowledge can be explained: the assumption that, as it can an eye or a wing, evolution is able to craft strong innate constraints on cognition. In this thesis, I re-examine that assumption with respect to a broad class of behaviours: those that we learn from other individuals.

The paradigm example from this class of behaviours is human language: our species-unique capacity for language rests upon the ability to induce complex, arbitrarily structured mental representations on the basis of inference from sparse, noisy behaviour observed in other individuals. There is widespread conviction that this ability depends upon inbuilt language-specific knowledge; that our inferences about language are guided by strong innate inductive biases. This view is motivated by two major observations: as a population, we tend to converge on a small subset of possible behaviours; as individual learners, we seem willing to draw conclusions that go beyond what is licensed by experience. As such, language has become the flagship case for nativism, and for the underlying principle that evolutionary reasoning legitimates that perspective. This thesis questions that status on the grounds that language, like many other human behaviours, is culturally transmitted. At its core is a general hypothesis: that cultural transmission radically re-structures the evolutionary process underpinning the formation of innate constraints on cognition. Throughout, I take the evolution of innate language-specific knowledge as a test case for this hypothesis, and ask how this process can inform our expectations about cognition.

In chapter one, I review the concept of nativism, its application to language, and its marriage to evolutionary reasoning. I argue that culture significantly complicates the evolutionary defence of nativism: where behaviours are transmitted culturally, the co-
evolutionary process underpinning the formation of innate knowledge includes complex properties that are not accounted for by existing approaches to the topic. Chapter two presents a review of existing co-evolutionary models: in addition to models of language, it covers a range of related topics including birdsong, sign languages, and social learning. Though many of these models are instructive, none comprehensively captures the co-evolutionary process that is induced by cultural transmission and underpins the formation of innate constraints on cognition for behaviours like language. Chapters three and four are studies in the dynamics of this process: there I derive and analyse a series of mathematical models, built around Bayesian models of cognition and iterated learning, that ask how culture shapes the evolution of cognition. On the basis of these models, I argue that culture can intervene in the formation of innate knowledge: the evolution of innate constraints on cognition for cultural behaviours is not straightforwardly comparable to a-cultural equivalents. However, rather than prohibit cognitive specialisation, culture can in fact facilitate rapid cognitive adaptation. By giving rise to rich co-evolutionary dynamics, culture can solve the problems it creates. Furthermore, the solutions which emerge are counter-intuitive: it allows small, potentially undetectable cognitive changes to have large population-level effects, yet rules out the evolution of strong innate constraints on cognition. In chapter five, I ask how the predictions of these models might be tested empirically, and consider their implications for experimental approaches to language-related cognition.
Chapter 1

Nativism, Evolution, and Culture

Nativism is the thesis of inborn knowledge: a philosophical doctrine concerning the hypothesis that knowledge is native to an organism. Nativism is linked historically and theoretically to the opposing doctrine of empiricism: the thesis that knowledge is acquired empirically through experience. The opposition of these doctrines has, for many centuries, traced deep divisions in philosophical and scientific inquiry into human nature. That these ideas have proven so persistent reflects a simple but pernicious tension. On the one hand, these doctrines are thought to imply broad and profound consequences; at one time or another, each has been closely associated with huge philosophical, political, moral, epistemological, scientific, and even metaphysical theses. On the other hand, direct evidence that favours or falsifies either doctrine is inherently difficult to obtain: though we have recently begun to try, looking inside peoples’ heads (and making sense of the results) to settle the argument has always been difficult. For this reason, debates surrounding these doctrines historically turned on logical reasoning, and more recently on reasoning from indirect evidence about their plausibility: in particular, both historically and in contemporary inquiry, the credibility of nativist philosophies has been fundamentally tied to theories concerning the origins of inborn knowledge: first meta-physical, now evolutionary.

In this chapter, I’ll present the case that nativism as a doctrine is rife, and that its credibility is inherently dependent on evolutionary considerations. My goal is to argue that this marriage of nativism and evolutionary reasoning is compromised by culture; that the human capacity to transmit knowledge culturally radically restructures the evolutionary model generally thought to support nativism. I’ll briefly review the
history of nativism (section 1.1), then discuss its contemporary standing (section 1.2). Section 1.3 covers the relationship between nativism and evolution, and sets out how culture can intervene. I’ll conclude by arguing that the flagship case for nativism in contemporary psychological science, linguistic nativism, is associated with an evolutionary rationale that, in the light of culture, requires revision.

1.1 Nativism: an enduring thesis

Accessible introductions to and histories of the nativism-empiricism dialogue can be found in Diamond (1974); Samet (1999); Stich (1975); Cowie (1999); Carruthers (1992); Carruthers et al. (2005); and Samet (2008). In particular, Stich (1975) provides a thorough historical overview, while Cowie (1999), Piattelli-Palmarini (2000), Wagner (2001), and Fodor (2001) provide modern readings of the historical debate. Here I present a brief overview of this history, since contemporary debates are clearest in this light, and because several of the core arguments I will make in this thesis have ancestry in this literature.

While some see the seeds of the debate in pre-history, specifically in the works of pre-Socratic philosophers Empedocles and Anaxagoras (e.g. Samet, 2008), the first known explicit nativist proposal, and the touchstone for nativist pedigree, appears in the works of Plato.

1.1.1 Platonic Nativism

The nativist hypothesis is first put forward in the Meno (Plato, 1961, trans.), a dialogue in which Socrates directs questions of geometry to a local slave. Socrates first demonstrates the slave’s ignorance of geometric theory, then poses a series of questions that each induce in the slave some form of introspective enquiry. When the questioning is complete, Socrates demonstrates that the slave has attained geometrical knowledge through introspection alone. The dialogue illustrates the doctrine of anamnesis: the idea that all knowledge is attained by recollection rather than empirical discovery; all learning is recall. Though the theory is hugely underspecified, and has since been almost universally rejected (Carruthers, 1992), it is a clear (and extreme) rejection of the empiricist premise (knowledge is acquired through experience) and articulation of a nativist hypothesis: knowledge is inborn.
Plato develops arguments in favour of the theory in a later dialogue, *Phaedo* (Plato, 1961, trans.). As *Meno* is the touchstone for the nativist hypothesis, so *Phaedo* is the touchstone for the line of reasoning that has come to be known as the *argument from the poverty of the stimulus* (APS). Here Plato illustrates the hypothesis that some forms of knowledge logically cannot be gained through experience. With reference to a postulated realm of Forms - abstract necessary or timeless immutable universals such as equality, beauty, and logic - Socrates reasons in this dialogue that because material experience is characterised by constant change and imperfection, knowledge of the Forms cannot be based on exposure to their material instantiation, and must therefore be pre-given. In addition to the content of the APS, this approach anticipates much subsequent nativist theorising in at least two further respects: 1) methodologically, by the use of indirect evidence, and; 2) philosophically, by equating a rejection of empiricism with support for nativism.

To understand why these ideas became so influential, and the gravity of the debates that followed, it is helpful to consider the implications of the theory as historically conceived. In addition to the psychological commitments of the thesis, Plato saw huge philosophical, political, and metaphysical consequences too. For example, it is possible to interpret Plato’s ideas as supreme justification for Socratic philosophical inquiry in civilisation, and as proof of the immortality of the soul. To summarise this reading: the anamnesis thesis solves the *paradox of inquiry* - the idea that it is useless to examine topics whose nature we do not already understand, because our ignorance prevents useful evaluation of the outcome of such inquiries - by cashing out inquiry in recollection. By appeal to knowledge that has been acquired outwith the lifetime of the individual, it also postulates a knowledge-gathering soul that existed before birth, and will continue to exist indefinitely after death.

In brief, Platonic nativism holds that all knowledge is innately present in us, waiting to be recollected, by deed of the history of an immortal soul. Subsequent debates have challenged and developed this view on many fronts, aiming either to construct new brands of nativism, or to discredit the nativist hypothesis entirely. The former enterprise was foundational to a school of thought born in Europe during the 17th and 18th centuries, known loosely as *continental rationalism*.
1.1.2 Continental Rationalism

While still implicitly functioning as a psychological doctrine, nativism during the 17th and 18th centuries was arguably centre stage for a second reason: it played a fundamental role in rationalist epistemology. Broadly speaking, the notion that some principle of knowledge or belief is innate was widely invoked to justify our assent to that principle: being innate somehow guarantees the truth or morality of an idea. This assumption was deployed widely in support of various epistemological agendas. For example, nativism was embraced by a group of philosophers working in England during the 17th century now known collectively as the Cambridge Platonists (Yolton, 1956). Cowie (1999), citing Yolton (1956), reads the Cambridge Platonists’ appeal to nativism as a theologically motivated defense against the contemporary spread of scepticism threatening Christian values: if Christian moral sensibilities are innate, they cannot be abandoned or questioned. Like Plato, nativists of this flavour envisaged innate knowledge that is fixed and eternally present, reflecting divine principles known to the soul.

However, the most influential nativist thinkers of the time were René Descartes (Descartes, 1647, trans.) and Gottfried Wilhelm Leibniz (Leibniz, 1704, trans.). Current scholars debate whether Descartes held all ideas to be innate, or just special instances, such as God, mathematics, truths, and natures (e.g. Gorham, 2002), and whether Cartesian innateness is a unified or multi-faceted concept (e.g. Boyle, 2009). Nevertheless, Descartes’s appeal to innate knowledge, however construed, seems to have served several higher philosophical ends: 1) to solve specific APSs; 2) to explain and justify belief systems, and; 3) to defend the existence of God. With respect to 1), innateness was the backbone of Descartes’s theory of perception and intelligibility. For humans to arrive at rich conceptual representations from experience with impoverished raw sensory stimuli, we must also be endowed with innate abstract concepts that permit recognition and interpretation:

"If we bear well in mind the scope of our senses, and what it is exactly that reaches our faculty of thinking by way of them, we must admit that in no case are the ideas of them presented to us by the senses just as we form them in our thinking. So much so that there is nothing in our ideas which is not innate to the mind or the faculty of thinking..." (Descartes, in Cottingham et al., 1985, p. 304)

Descartes’s reasoning in support of 2) engenders two subtly distinct suggestions.
Firstly, Descartes proposes that innate ideas are fixed in our belief system: that is, we cannot disbelieve an innate idea. Though this is the weaker of the two propositions, it suggests a controversial conception of innateness that generated a long-standing debate, to which we will turn in section 1.1.3 and often thereafter in this thesis. The stronger proposition holds that innateness implies justification, or truth: an idea being innate is reason for believing we are right to assent to it, or that it is true. Cowie (1999, p. 10-12) provides a thorough discussion of these proposals, concluding that innateness in fact does very little explanatory work in Descartes’ epistemology: rather, it is the appeal to divine intervention that justifies belief, and for Descartes, innateness implies Godly origin; in the absence of alternate theories of design, innate ideas can only have been put there by God, and are so guaranteed. Nativism and Godly metaphysics are tightly coupled in this period, particularly for Descartes and followers, as noted in 3) above. Though the details of Descartes’ proof for the existence of God are unimportant for this review, the coupling of nativism and religion will become highly significant.

The other prominent nativist thinker of the time is Leibniz. Leibnizian nativism is a descendant of Cartesian nativism. That is, Leibniz too defended the idea of innate concepts and their implications, but tempered Descartes’ nativism in response to prevailing debates. In particular, Leibniz was forced to defend nativism against empiricists who challenged the contemporary understanding of innateness. It is in this period that the nativist position begins to fracture into varieties, as a direct result of debates concerning the nature of innateness. Though it is fought on quite different terms now, this debate continues today. Then as now, it attracts extreme reactions, from those who see the innateness debates as one of the most enduring engines for philosophical and psychological research in the history of these disciplines (Plotkin, 2007; Spelke, 2010), to those who see the concept as useless and call for its retirement (Spencer et al., 2009).

Since Leibnizian nativism is tightly coupled with empiricist objections to innateness, it is worth introducing this debate first, and revisiting Leibniz in that light.

### 1.1.3 Empiricism & the Enlightenment

Though ingredients of empiricism can be traced back at least as far as Aristotle, its major role in the history of nativism lies in the Enlightenment, and more specifically with a 17th/18th century movement now known as British Empiricism. The movement
is generally thought to have gathered momentum with the work of arch empiricist John Locke, an English Enlightenment philosopher. Broadly, the Enlightenment period engendered an intellectual backlash against religion and tradition as credible sources of knowledge, in favour of empirical and scientific inquiry. This wider philosophical and scientific context is crucial to an understanding of the innateness debates since, as discussed in §1.1.2, nativism was at this point closely tied to traditional metaphysical and epistemological commitments. In 1689 Locke published the hugely influential *An Essay Concerning Human Understanding* (Locke, 1689), which included a lengthy critique of nativism, and a strong challenge to the notion of innateness in particular. First under question is an influential line of reasoning, common to nativist accounts of the time and still widely invoked by contemporary nativists, which Locke called the *argument from universal consent*: the idea that being universal among the species implies some principle of knowledge that it is innate. Locke rejects this inference at the outset of his *Essay*, claiming that:

"This argument, drawn from universal consent, has this misfortune in it, that if it were true in matter of fact, that there were certain truths wherein all mankind agreed, it would not prove them innate, if there can be any other way shown, how men may come to that universal agreement in the things they do consent in; which I presume can be done." (Locke, 1689, p. 12)

However, despite challenging the logic of this reasoning, Locke goes on to take it at face value and instead challenge universality, both in general and with respect to specific aspects of knowledge. By doing so he brings the notion of innateness itself into focus, chiefly by questioning what might nowadays be termed the *ontogenetic status* of innate ideas. In simple terms, Locke claims that, for universal consent to legitimately indicate innateness, there can be no exceptions. Since it is easy to find "idiots" and children (these are Locke’s reference points, not mine!) who are entirely unaware of supposedly innate aspects of knowledge (logical, mathematical, metaphysical, and religious principles, for example), such principles cannot be innate. This reasoning assumes of the nativist an extreme conception of innateness: that innate ideas are necessarily known and consented to consciously from birth. While to the contemporary reader this may seem an incredible thesis, it was at the time a fair reading of nativist rhetoric (recall for instance Descartes’ suggestion that we cannot but assent to innate
ideas, discussed in section 1.1.2). Though some have accused Locke of arguing against
a straw man here, his reasoning drew attention to the fact that, to be a viable explana-
tory concept, innateness must mean something more intricate than present at birth. Cru-
ially, Locke also permits the nativist a weaker notion he terms dispositional innateness:
knowledge is innate if we are disposed to acquire it. Considering various mechanisms
that might implement such dispositional coming-to knowledge (e.g. exposure to that
knowledge, the use of reason, etc.), he concludes that none are significantly distinct
from the empiricist acquisition proposal, and so declares nativism redundant.

Locke’s reflections on innateness forced a re-formulation, or at least re-articulation,
of the nativist position. Chief among Locke’s challengers is W. G. Leibniz (introduced in
section 1.1.2), a rationalist who re-formulated Cartesian nativism in the light of Locke’s
objections. Leibniz made explicit in nativist theories the notion of the cognitive un-
conscious: ideas can be present in the mind but not always consciously accessible. He
proposed an analogy for his brand of dispositional innateness: in place of Locke’s tab-
ula rasa, Leibniz compared the newborn mind to a slab of marble, enriched with veins,
or weak points, that trace the structure of a statue. The statue is implicit in the marble
slab, realised only through experience, but not visible to inspection.

Thus began the innateness debates that continue today. We can see in the Locke-Leibniz
exchange the earliest models for the combination of experience and innate predisposi-
tions in knowledge acquisition. It is clear from this point onwards the the significant
differences between nativist and empiricist accounts of knowledge acquisition map on
to the details of psychological or epistemological proposals, rather than nature and nur-
ture per se. Historically, empiricism dominated rationalism for centuries post-Locke,
being the backbone of Enlightenment thinking, and gaining support from numerous
influential philosophers such as Hume, Berkley, Stuart-Mill, and Quine. Though the
causes of nativism’s decline in this period are undoubtedly many and varied, two
points in particular are relevant here. First, nativism is closely tied with rationalism:
tied conceptually, since both make appeal to mind internal causes of knowledge, and
tied historically, since prominent nativists were also rationalists. As empiricism re-
placed rationalism as a general world-view, rationalists’ nativist commitments were
tarred with the same brush as their metaphysical commitments, being “backward and
unscientific” (Samet, 2008) by association. The second point is closely related, but per-
haps more fundamental. Pinpointing an explanation for the decline (and subsequent re-
surgence, see §1.2) of nativist theorising, Carruthers (1992) suggests that nativism was
more than just casually associated with unfashionable metaphysical doctrines: it was,
at that time, inherently dependent upon them. More precisely, there was at the time
no natural explanation for how innate knowledge could come to exist; pre-Darwin/
Wallace, nativism was deeply unappealing to the Enlightenment thinker because the
only hypothetical explanation for the presence of innate knowledge was supernatural:
recourse to nativism was recourse to God.

In summary, this review has revealed the seeds of several ideas that will feature heavily
in this thesis: nativism is only credible where there exists some plausible explanation
for the presence of innate ideas; being universal does not necessarily imply of a trait that
it is innate; and innateness does not necessarily imply fixed inflexibility or inevitability,
but rather reflects a spectrum of dispositional possibilities.

1.2 Contemporary Nativism

Nativism resurfaced, after centuries of empiricist reign, during the second half of the
20th century. Contemporary nativism is for the most part now divorced from the many
metaphysical and epistemological associations so rife during earlier periods, and in-
stead functions principally as a purely psychological doctrine. Excellent overviews of
contemporary nativism en général can be found in Carruthers et al. (2005), Carruthers
et al. (2007), and Carruthers et al. (2008). Those volumes contain representative per-
spectives on contemporary nativism in a range of psychological domains, including vi-
sual perception (e.g. Scholl, 2005), spatial reasoning (e.g. Shusterman and Spelke, 2005),
mathematical cognition (e.g. Laurence, 2005; Leslie et al., 2008; Laurence and Margol-
isis, 2008), social reasoning (e.g. Povinelli and Prince, 2005; Wynn, 2008), intentionality
understanding (e.g. Johnson, 2005), homicidal instinct (e.g. Duntley and Buss, 2005),
morality (e.g. Greene, 2005; Wynn, 2008), religious belief (e.g. Atran, 2007), folk psy-
chology (e.g. Segal, 2008), and judgement and decision-making (e.g. Santos and Laksh-
minarayanan, 2008). Other psychological domains that enjoy notable nativist perspec-
tives include: concepts (e.g. Cain, 2004), logic (e.g. Crain and Khlentzos, 2007), music
perception (e.g. Justus and Hutsler, 2005), folk physics (e.g. Carey and Spelke, 1994),
folk biology (e.g. Atran, 1990), abstract number comprehension (e.g. Izard et al., 2009),
mate choice preferences (e.g. Todd et al., 2007), cooperation and social motivation systems (e.g. Delton and Sell, 2014), conflict resolution instincts (e.g. Sell et al., 2009), visual attention (e.g. Ro et al., 2001), navigational abilities (e.g. Krasnow et al., 2011), and kin detection (e.g. Lieberman et al., 2007), among others. Nativism is represented throughout the psychological sciences: it is alive and kicking, and in the eyes of many has a strong future ahead (Margolis and Laurence, 2012; Spelke and Kinzler, 2009). This is not to say that nativism is dominant: whether or not that is true is difficult to establish with any confidence. While it is easy to find broad claims of nativism’s prominence, empiricism is still thought to be the default position from which to approach cognition:

"Nativist theorising is thriving. Nativism is once more at the forefront of contemporary developmental and cognitive theory." (Carruthers et al., 2005, pp.1)

"Empiricism is often taken to be the default view both in philosophy and in cognitive science." (Margolis and Laurence, 2012, pp. 693)

However, what is certain is that, throughout the cognitive sciences, there are vocal proponents and critics of the nativist approach to cognition: nativism is a prominent and once-again credible psychological perspective:

"Over the last fifty years, the tide has turned against [empiricism]...It has been supplanted, for many psychologists, by the view that perception, thought, values, and actions depend on domain-specific cognitive systems....Each system has its own innate foundations and evolutionary history, and each system functions to treat specific kinds of information for specific purposes. The best known example is the language faculty: an innate, special-purpose system with a distinctive internal mode of operation and a distinctive developmental course." (Spelke, 2010, p. 204)

In light of nativism’s rocky history, two questions immediately arise: i) How did nativism regain prominence and credibility? and; ii) What does the 21st century nativist believe? The quote from Spelke above hints at answers to both. The credibility of nativist philosophies, and the nature of their core propositions, are both inextricably linked to biological evolution. By providing a mechanism for the formation of innate knowledge, evolution protects nativist philosophies against the inevitable regress to
Godly intervention that shut down their development in the 17th/18th century (Carruthers, 1992). The prominence of nativist theories is widely attributed to the huge success and influence of the nativist enterprise for the signature human cognitive capacity: language. In the following sections (and throughout this thesis), I’ll take linguistic nativism to be the paradigm example of modern nativist theorising, and concentrate on this topic. Though inevitably some of what follows will reflect narrowly linguistic concerns, much will echo issues that arise throughout contemporary nativism.

1.2.1 Linguistic Nativism

Human language is the flagship case for nativism. Language is often assumed to provide the strongest evidence that the human mind possesses substantial innate structure (e.g. Baker, 2005). The success of nativist enterprises in linguistics is taken as justification for nativism in other domains (Pinker, 1994; Tooby and Cosmides, 1992), and has inspired researchers in other disciplines to borrow arguments and conclusions as models of inquiry to support nativist theories about other domains (Hirschfeld and Gelman, 1994), such as moral psychology (e.g. Dwyer, 2007; Dwyer et al., 2009; Hauser, 2008).

The resurgence of nativism in linguistics, often dubbed the beginning of the cognitive revolution, is widely agreed to have begun in the mid twentieth century, with Chomsky’s famous rejection (Chomsky, 1959) of the then prominent empiricist paradigm behaviourism. Certainly, Chomskyan linguistics has traditionally been the bedrock of linguistic nativism. A central tenet of that enterprise is the hypothesis that language acquisition and use is facilitated by a species-unique innate set of special cognitive capacities (see Clark and Lappin, 2010, for a concise overview of this perspective), often referred to collectively as Universal Grammar (UG) (Chomsky, 1965, 1987, 1995). The hypothesis is traditionally defended along two independent lines of reasoning, both familiar from section 1.1: the argument from the poverty of the stimulus (APS), and the argument from universal tendencies (AUT). The APS has arguably been primary, but is nevertheless extremely controversial (Cowie, 1999; Collins, 2003; Pullum and Scholz, 2002). Its history (Thomas, 2002), contemporary standing (see Pullum and Scholz, 2002, and papers in that special issue), and its overall significance to linguistic nativism (Clark and Lappin, 2010, chapter 2) are well documented (see Piattelli-Palmarini and Berwick, 2013, for a range of recent perspectives on the APS). In essence, it remains unchanged.

\[\text{1}\]See Hinzen (2014) for a brief history of the concept of Universal Grammar.
from Plato’s formulation (section 1.1.1), but applied to language: observable linguistic
data are sparse, noisy and otherwise so impoverished that a learner without appropri-
ate inborn dispositions could not acquire the generating grammar as the human child
does, therefore, the child must possess substantial innate domain-specific biases\footnote{The reader might argue that the clause following ‘therefore’ is not implied by the preceding clause. See Laurence and Margolis (2001) and Pullum and Scholz (2002) for exemplary deductive formulations of the linguistic APS, and Cowie (1999) for thorough discussion of this observation.}. The
final clause appeals to notions of innateness, domain-specificity, and inductive bias:
these concepts are central to contemporary nativist theorising, and I will discuss each
in depth shortly.

The AUT has also been foundational, but in a slightly different way. Where the APS
is taken to represent \textit{a priori} motivation for the existence of innate constraints, universal
linguistic tendencies represent \textit{supporting evidence}:

"Striking asymmetries in the frequencies with which different linguistic pat-
tterns occur in the world’s languages constitute a central body of evidence
for the cognitive biases operative in language acquisition" (Culbertson et al.,
2013, pp. 393)

Species-wide shared innate constraints predict regularity in the typological distri-
bution of linguistic structures by delineating a set of possible languages; they dictate
that all languages should obey those constraints. As a consequence, a substantial com-
ponent of the nativist enterprise in linguistics has been the search for universals of
language: typological regularities that reflect linguistic properties common to all lan-
guages:

There has also been very productive study of generalizations that are...directly
observable: generalizations about the word orders we actually see...These
universals are probably descriptive generalizations that should be derived
from principles of UG. A major research project within generative gram-
mar has been to explain them in such terms. (Chomsky, in Cela-Conde and
Marty, 2002, pp. 33-34)

A thorough overview of contemporary perspectives on the status of language uni-
versals can be found in Christiansen et al. (2009). While many have questioned the exis-
tence of significant linguistic universals (e.g. Evans and Levinson, 2009), or questioned
the AUT line of reasoning directly (Newmeyer, 2007), the basic assumption that typological regularities reflect, and imply, underlying specialised cognitive constraints remains central to nativist theories of language (e.g. Pinker and Jackendoff, 2009a; Prince and Smolensky, 2004).

Given these foundations, and the broad influence of linguistic nativism, how well accepted are nativist claims about language? To be sure, there is a wide consensus that the human capacity for language relies upon innate mechanisms:

"In point of fact, there is general agreement that the acquisition of language is innately guided - this much has been widely acknowledged even by those opposed to the idea..." (O'Grady, 2008, pp.620)

"Clearly human beings have an innate, genetically specified cognitive endowment that allows them to acquire natural language." (Clark and Lappin, 2010, pp.1)

"Today, there is a near-complete consensus that the human capacity to acquire language has a strong genetic basis, aspects of which evolved uniquely in our species. From this viewpoint, everyone is a nativist." (Fitch, 2011a, pp. 325)

"What I think we can say without question is that the child does not come to the task of language acquisition empty headed." (Kirby, 2004, pp. 928)

"the existence of [innate] constraints [that facilitate language acquisition] cannot be seriously doubted." (Hauser et al., 2002, pp. 1577)

So are all linguists nativists? Not in the slightest. The excerpts above are accurate nevertheless because contemporary nativism denotes more than a belief in the presence of innate machinery. And contemporary empiricism denotes more than a denial of innate machinery. The terms of the debate have changed:

"At some level, of course, we all concur in the existence of some degree of innate specification. The difference in positions concerns how rich and domain-specific the innately specified component is..." (Karmiloff-Smith, 1998, pp. 389)
"The debate between opposing perspectives does not concern the existence of innately specified cognitive capacities....What remains in dispute is the nature of this innate ability, and, above all, the extent to which it is a domain-specific linguistic device." (Clark and Lappin, 2010, pp. 1)

As these excerpts suggest, the debate between nativists and empiricists is now, and perhaps has always been (Cowie, 1999), more subtle than presence vs. absence of hard-wired knowledge. All contemporary theories of cognition assume some degree of innate knowledge. Though at the time of writing the sentiment was far from unanimous, the following quote from Popper is an early articulation of the current consensus:

"...[the] tabula rasa theory is absurd: at every stage of the evolution of life and of the development of an organism, we have to assume the existence of some knowledge in the form of dispositions and expectations" (Popper, 1972, pp. 71)

Though the substance (and value) of contemporary nativist-empiricist debates has itself been the subject of discussion (e.g. Spencer et al., 2009; Spelke and Kinzler, 2009; Samuels, 2002), and mis-attribution of argument between camps is a common theme (see e.g. Cowie, 1999; Simpson, 2005, for discussions), there are at least two issues that represent clear lines of division between the philosophies. Naturally, since both philosophies involve commitment to many subtle hypotheses that may not be deterministically dependent upon each other, the following should be taken as strong tendencies, rather than strict requirements. First, the nativist, but not the empiricist, hypothesises the existence of innate mechanisms that are specialised for the domain of language; contemporary nativism entails support for domain-specificity. Second, the nativist approach admits a greater strength or magnitude of innate machinery; where the empiricist hypothesises innate machinery that is weaker in influence, or slighter in volume, the nativist supposes a primary role for innate machinery that strongly constrains human language.

---

3 Though see Clark and Lappin (2010) for an alternative perspective on the relationship between historical and contemporary debates.

4 For example, the nativist might hypothesise domain-specific strong constraints on structural form, but nevertheless allow a role for domain-general statistical inference procedures. Likewise the empiricist might deny domain-specific idiosyncratic computational constraints, but allow for domain-specific adaptations in the auditory cortex. I will revisit the question of multiple commitments later in the chapter, but note here that my characterisation of the debate will focus on prototype tendencies, which should be read as such.
What does it mean to say that a capacity is domain-specific? While often tightly coupled with evolutionary history (Hirschfeld and Gelman, 1994) and innateness (Khalidi, 2001), the idea is not intrinsically connected to either of these. To be domain-specific simply means that the capacity operates on input from a specific domain: a language-specific cognitive capacity is dedicated to language. Khalidi (2001) gives a good working definition:

"To say that a cognitive capacity or set of beliefs or collection of ideas is domain-specific is to say that it is dedicated to solving a restricted class of problems relating to a certain field of inquiry or range of phenomena" (Khalidi, 2001, pp. 193-194)

Whether cognitive development proceeds from domain-general or domain-specific capacities is a crucial dividing line, and a fundamental issue for nativist theorising in numerous psychological domains (see e.g. Sloutsky, 2010, and papers in that special issue).

What does it mean to claim that innate cognitive machinery is large in magnitude or strongly constraining? Crucially, this is not a tautology: as noted in section 1.1, innateness does not imply hard-wired knowledge; innate dispositions can equally be weak or strong. It is worth noting at this point some difficulties with the concept of innateness itself. Despite widespread usage, the notion of innateness has received significant criticism. Innateness is a controversial concept, historically and in contemporary discussions (see e.g. Samuels, 2004; Carruthers et al., 2005; Mameli and Bateson, 2011; Griffiths, 2002). Discontent with the concept has a long pedigree, dating back to early critiques of the concept of ‘instinct’ in biology (e.g. Lehrman, 1953). The force of the objection to the term generally lies in its polysemy: throughout the biological and cognitive sciences, it has been used to refer to a number of related but distinct ideas, such as presence at birth, genetic heritability, canalisation, and so on. Some have suggested the term be replaced: for instance, Shea (2012) suggest "inherited representation" would be a more concrete concept. While the resolution of these issues is important, I don’t intend to get bogged down in them here. My view, following others, is that terminological ambiguity does not prevent productive inquiry: "Scientific progress in investigating a kind does not generally depend on having an airtight characterisation of that kind" (Carruthers et al., 2005, p. 4). While terminological complications can be disruptive for
certain kinds of enquiry, black-boxing the details of innateness can be perfectly acceptable in others (Fitch, 2012). Where I use the term, it should be understood in roughly the manner described by Carruthers et al.’s working definition:

"... a cognitive mechanism, representation, bias, or connection ... [is] innate to the extent that it emerges at some point in the course of normal development but is not a product of learning.\)"(Carruthers et al., 2005, pp. 1)

Rather than restrict innateness to strict notions such as presence at birth, I’ll assume a concept more like dispositional innateness (Stich, 1975; Khalidi, 2007), known since Locke (see section 1.1.3), which allows that learners can be innately predisposed, to an arbitrarily weak or strong extent, toward certain traits. I’ll argue shortly (in section 1.2.1.2) that the clearest way to understand how innate knowledge can be strongly constraining is to couch the concept in terms of inductive bias, and that there now exist computational models of cognition that accommodate this notion. To be absolutely clear: nativism does not equal a belief in innateness \textit{per se}. Rather, it equates to a belief in (domain-specific) machinery that is innate and strongly constraining.

1.2.1.1 Nativist Linguistic Theories

Nativist approaches to language are popular (e.g. Crain and Pietroski, 2001; Pinker and Jackendoff, 2009a; Gómez et al., 2014; Culbertson and Adger, 2014; Crain, 1991; Prince and Smolensky, 2004), and take many forms. Though there is huge variation in the extent and content of innate machinery hypothesised to support language (see Tomasello, 2007), nativist theories are connected by a common commitment: each supposes that humans are equipped with a suite of language-specific genetically-specified cognitive mechanisms that impose strong constraints on language acquisition. The set of related theories that appeal to a rich UG constitute prototypical nativist theories. For instance, the earliest proposals in generative grammar (e.g. Chomsky, 1965) hypothesise the existence of an innate domain-specific cognitive module that supports language acquisition. That module (UG) receives primary linguistic data as input; it imposes constraints on the set of generative grammars the learner can hypothesise; it determines the nature of the structural descriptions the learner can assign to an observed string; it encodes knowledge of the likelihoods of observing particular strings from particular grammars,
and; it specifies a language-specific evaluation metric that favours certain inferences over others.

Later proposals include the Principles and Parameters framework (e.g. Chomsky, 1981), which has been called "the bona fide theory of innateness" (Yang, 2004, pp. 451). This theory is based on a hypothesised language-specific device (UG) that constrains the set of grammars a learner can acquire to be finite. It achieves this by encoding innate preferences that favour certain linguistic features over others: some linguistic features are imposed by inviolable principles that prohibit the acquisition of any language that does not accord with these principles, others are influenced by parameters that allow a constrained set of possible forms to be chosen among.

Each of these basic characterisations of UG persists in modern linguistic theory. For instance, Culicover and Jackendoff (2005) argue for a version of generative grammar they call Simpler Syntax, which removes much theory-internal complexity from earlier versions of generative syntactic theory, but maintains the core assumption that there exists a domain-specific innate cognitive module that favours the acquisition of languages with certain properties over others. UG under this conception is hypothesised to specify: principles that govern the structure of phrases, such as the X-bar theory, principles of agreement and case-marking, and constraints governing long-distance dependencies, alongside; principles that govern the syntax-semantics interface, such as constraints on mappings between thematic and syntactic roles; proto-linguistic structural principles, and; functional principles, such as constraints that favour harmony in the ordering of heads and complements across phrasal types (see Culicover and Jackendoff, 2005, chapter 1 for a concise summary). Similar conceptions of UG can be found in Jackendoff (2002); Pinker and Jackendoff (2009a); Jackendoff and Pinker (2005); Pinker and Jackendoff (2005).

This basic approach, which assumes rich, extensive innate specification of specific linguistic properties, thereby constraining learners to acquire only those linguistic systems that possess these qualities, is reasonably common. For example, Wunderlich (2004) argues that UG provides domain-specific specialisations that favour: robust categorical distinctive features for linguistic units; a basic distinction between predication and reference; distinct lexical categories, including noun and verb; hierarchical organisation of arguments and the consequent asymmetry between subject and object; and constraints on reference tracking.
Likewise, contemporary theories in the Principles and Parameters vein (e.g. Baker, 2001; Yang, 2004) also hypothesise a UG that includes rich innate language-specific structure. Specific proposals for innate constraints include: principle C of the binding theory, which is thought to impose constraints on various phenomena such as anaphoric relations and scope interactions (Crain and Pietroski, 2006); an inviolable constraint that enforces structure-dependence (Crain et al., 2010; Crain, 1991); island constraints (Crain et al., 2010); and downward entailment rules, which impose constraints on phenomena such as disjunction and negation (Crain et al., 2010; Crain and Pietroski, 2006).

Though nativist theories of language are typically associated with the generative grammar movement in syntax, other corners of linguistics have produced nativism too. For example, "the presence of biological linguistic constraints on language acquisition" (Gómez et al., 2014, pp. 5837) is often invoked to explain structure in phonology: in relation to the Sonority Sequencing Principle, Gómez et al. (2014) claim that "humans possess early, experience-independent linguistic biases concerning syllable structure that shape language perception and acquisition" (Gómez et al., 2014, pp. 5840). Similarly, on the basis of widely dispreferred syllable onset consonant clusters, Berent et al. (2008) "conclude that language universals are neither relics of language change nor are they artifacts of generic limitations on auditory perception and motor control - they reflect universal linguistic knowledge, active in speakers’ brains" (Berent et al., 2008, pp. 5321). More generally, Optimality Theory (Prince and Smolensky, 2004) attributes to an innate domain-specific UG a collection of primitives: universal constraints whose interaction determines linguistic form. It finds most application in phonology (see e.g. McCarthy, 2004), but has also been applied to other domains such as syntax (see e.g. Legendre et al., 2001) and morphology (see e.g. Xu, 2011). The set of constraints hypothesised varies across domains, and there is debate over which constraints should be classed as domain-specific (Haspelmath, 1999). As an example, though, Culbertson et al. (2013) discuss a set of possible primitives that constrain nominal word-order. These include HEAD-L and HEAD-R, which respectively enforce ordering of heads to the left or right of their complements, and NUM-L, which constrains numeral heads to be positioned left of their complements. The set of grammars available to the learner is constrained by the possible rank permutations of these constraints.

All of the theories above are classically nativist: they assume rich, strongly con-
straining domain-specific cognitive mechanisms that directly determine much of the structure present in language\(^5\). In Clark and Lappin’s (2010) terminology, we might call this class of approaches the set of *strong bias* theories. Interestingly, the contemporary Chomskyan school of thought advocates theory that, on the terms of this discussion, is not classically nativist at all. The minimalist program (Chomsky, 1995) is devoted to the expulsion of rich specification from syntactic theory, and consequently also from the associated schema for UG. Much of the computational power and idiosyncrasy previously attributed to innate domain-specific rich constraints is now attributed to a basic computational procedure known as *Merge*, alongside the so-called "third factor", which denotes cognition-external regularities (Chomsky, 2005) of this universe. An influential articulation of the new position concerning UG is given in Hauser et al. (2002). There Chomsky and colleagues hypothesise a very slim domain-specific innate endowment indeed: only the capacity for recursive computation, and its interfaces with other linguistic systems such as the lexicon, are claimed to be language-specific. Earlier, following virtually everybody else, I defined nativist theories as those which postulate innate machinery that is domain-specific and extensive. By both of these measures, minimalist approaches to language as currently articulated\(^6\) are not nativist theories.

### 1.2.1.2 A Richer Ontology

The case of minimalism is useful to this discussion in another respect: it highlights how the dimensions outlined, domain-specificity and volume of proposed innate machinery, define a range of possible theories. The first of these dimensions, domain-specificity, defines a clear division. The second of these dimensions has a number of possible interpretations. I favour the following, which will play a significant role throughout this thesis. We can interpret theories about the nature of language-related cognition as specifying the degree of *bias* present in the learner (e.g. Kirby, 2004). Rather than think in terms of cognitive modules, for instance, we can think in terms of innate biases (Cummins and Cummins, 2003) that can be arbitrarily weak or strong:

\(^5\)See Boeckx and Leivada (2014) for a concise summary of linguistic theories that appeal to strong innate knowledge.

\(^6\)It is a common sentiment to suppose that minimalist theories as currently conceived have insufficient explanatory power, and will eventually be forced to either: acknowledge much more intricate innate machinery than they do currently, or; develop the case for domain-general machinery and/or third factors, in which case they will be ontologically very similar to existing empiricist approaches to language (see e.g. Pinker and Jackendoff, 2005; Newmeyer, 1991, for relevant discussions).
"What are the capabilities and limitations of human language learning? According to classical arguments from linguistics and the theory of learning, answering this question involves discovering the biases of human language learners." (Culbertson et al., 2013, pp. 392)

Contemporary researchers in various domains, including syntax (e.g. Culbertson, 2012; Culbertson et al., 2012; Boeckx and Leivada, 2014) and phonology (e.g. Wilson, 2006; White, 2013), have defended the idea that constraints on language learning are well captured by the notion of inductive biases that can be strong and inviolable or weak and defeasible. Under this interpretation, traditional nativist theories of language hypothesise machinery that strongly biases the learner:

"We have suggested that UG may be a set of...strong biases on the child’s learning...". (Pinker and Jackendoff, 2009b, pp. 466)

Absolute or inviolable constraints, such as restrictions on hypotheses and evaluation criteria in early accounts of UG, principles in later conceptions, and primitives in optimality theory, represent the strongest possible inductive biases. However, biases can also be weak. This highlights the first major advantage of this perspective: it reveals a broader class of possible theories:

"...the space of possible ways in which knowledge might be innate in a system is very large and includes some very subtle cases." (Clarke, 1993, pp.180)

"The upshot is that we can now begin to explore a variety of ways in which knowledge may be partially innate ... we can think quite seriously in terms of "90% or 10% of an innate idea."" (Clarke, 1993, pp. 187)

Crucially, on these terms we can conceive of theories that assume a domain-specific but weakly biasing cognitive architecture. Proposals in this vein wouldn’t be nativist in the sense I have been assuming. The term minimal nativism has been used by several authors, quite independently it appears, to refer to theories of just this kind: in philosophy of mind by Ramsey et al. (1991) and by Clarke (1993), and in cognitive science by Plotkin (2007) and by Ullman et al. (2013). Clarke’s version describes the idea I am alluding to:
"Minimal Nativism, in my usage, does not signify any lack of task specificity in the biases. Rather, it signifies something about the way such task-specific biasing may be achieved. In particular, such biasing need not involve anything very rich, or complex, or rule-like. Small details concerning initial configurations." (Clarke, 1993, pp.186)

Though they are logical possibilities, theories that appeal to this kind of innate disposition immediately face an explanatory problem. Nativist theories that postulate strongly biasing architecture naturally predict linguistic universals, as discussed above. Weakly biasing cognitive architectures don’t straightforwardly enjoy this explanatory potential: they require some additional explanation of how weak biases could give rise to universals. For instance, Clark and Misya (2009) acknowledge this challenge, but suggest that "gently biased neural architecture can lead, in rich developmental and ecological context, to the reliable ... emergence of specific bodies of knowledge" (Clark and Misya, 2009, pp. 256, emphasis added). Thus, developmental and ecological context must shoulder some of the explanatory work. I’ll defer further discussion of this point until section 1.3.2, but note that much of this thesis will be based upon the idea that cultural transmission can play exactly that role.

In summary, the terms of contemporary dialogues about the nature of language-related cognition carve out a range of possible approaches: if we are arguing about the domain-specificity and strength of cognitive constraints, hard-line nativism and empiricism are not the only options. Table 1.1, which I have adapted from a similar proposal by Ullman et al. (2013), portrays a simplified schema of possible approaches, assuming binary dimensions for domain-specificity (domain-general vs. domain-specific) and constraint strength (weak vs. strong).

<table>
<thead>
<tr>
<th>Domain-specific</th>
<th>Weak</th>
<th>Strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimal Nativism</td>
<td>Classic Nativism</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.1: A simple schema to distinguish various approaches to the cognitive architecture supporting language, adapted from Ullman et al. (2013).

Classic nativist and empiricist proposals belong in the top right and bottom left coordinates respectively. The top left coordinate corresponds to the class of theories just described, which appeal to domain-specific but weakly biasing cognitive capacities: Clarke’s *minimal nativism*. The bottom right coordinate is a little like strong empiri-
cism: theories in that vein suppose strongly biasing or powerful learning mechanisms that function generally across domains. In fact this class of theories is fairly well re-presented in contemporary dialogues about language. Cognitive grammar and some varieties of construction grammar both attribute structure in language to innate cognitive capacities that are rich and powerful, but not domain-specific. More generally, it is fairly widely agreed that language learning involves general capacities of powerful statistical inference (e.g. Saffran, 2003), which seem to be robust, early-developing central features of wider cognition (e.g. Denison and Xu, 2013).

Crucially, the kinds of cognitive architecture alluded to in table 1.1 are not mutually exclusive (see e.g. Yang, 2004): it is perfectly possible, for instance, to hypothesise that learners are equipped with powerful capacities for statistical inference and strong innate domain-specific biases. Approaches of this kind, which borrow from various theoretical camps to build comprehensive and subtle pictures of cognition, are becoming extremely common. Computational advances have been central to this line of progress: for ends theoretical and practical, computational models of cognition have been developed that combine powerful statistical inference with rich, domain-specific representations (see e.g. Tenenbaum et al., 2011). At the forefront of this enterprise are Bayesian probabilistic models of cognition: computational models built upon the principles of rational inference. Bayesian models have been applied to a wide range of cognitive phenomena: good overviews of their application are provided by Chater et al. (2010) and Griffiths et al. (2010); gentle introductions to the foundations of these models are given in Perfors et al. (2011a), and in the special issue of Cognition headed by Chater et al. (2006). Probabilistic models have huge application to language acquisition (see e.g. Chater and Manning, 2006), and have been applied directly to questions of innateness and representation central to nativist reasoning about language, such as poverty of the stimulus and learnability arguments (see e.g. Hsu et al., 2011; Perfors et al., 2011b). Bayesian approaches to cognition offer a framework that can represent a wide range of the types of cognitive architecture described above. By combining general statistical inference procedures with potentially rich representations they offer a principled marriage of domain-general and domain-specific aspects of cognition. Arbitrary prior distributions provide a method for explicitly representing innate inductive biases, which captures the notion of dispositional innateness perfectly. Several authors have argued that these approaches are the future of the nativism-empiricism debates (Spelke and Kinzler, 2009;
Scholl, 2005): they represent a middle ground that allows principled combination of knowledge kinds, and precise quantitative formulations of cognitive theories (see e.g. Xu, 2008, for a discussion of this perspective, and a review of empirical evidence that supports it). The nativist, the empiricist, and most of those in-between can be explicitly positioned within this framework. Clark and Lappin (2010) summarise the point well:

"One way to understand the difference between domain-general learning and domain-specific learning is to consider an idealized form of learning. One of the most general such formulations is Bayesian learning...On this approach, we can achieve a precise characterisation of the contribution that domain knowledge makes, in the form of a prior probability distribution. In domain-specific learning, the prior distribution tightly restricts the learner....By contrast, in domain-general learning, the prior distribution is very general in character. It allows a wide range of possibilities... " (Clark and Lappin, 2010, pp. 3)

I will build much of the rest of this thesis on Bayesian cognitive models. However, for now, it is enough to conclude that contemporary theorising about the innate mind and language includes a spectrum of possibilities that extend the traditional nativist-empiricist positions, and that there exist formal and computational tools to accommodate the new playing field.

In summary, nativist theories are common in contemporary psychological science, but they are not the only game in town. In the next section, I discuss one of the major reasons why nativism has enjoyed such relative prominence. Nativist theories, but generally not their alternatives, enjoy (and depend upon) the support of a hugely influential line of reasoning: they are made credible and plausible by evolutionary theory.

1.3 Evolution & Nativism

Nativism and evolution are closely connected. Biological evolution provides the mechanism through which domain-specific innate knowledge can be formed: it rescued nativism from unpalatable metaphysical associations that compromised the philosophy historically (Carruthers, 1992). The application of evolutionary reasoning to questions of cognition has a long history (see Plotkin, 2004, 2007, for thorough overviews), and is the foundation of much contemporary nativist theorising. Evolutionary Psychology,
which aims to explain cognition with reference to its evolutionary history, is a burgeon-
ing enterprise, and has been called "the second wave of the cognitive revolution" (Cos-
mides and Tooby, 2013). At least among those who practise evolutionary approaches
to cognition, the perspective is thought to be inherently tied to nativism rather than
empiricism:

"Blank-slate theories of human intelligence propose that reasoning is carried
out by general-purpose operations applied uniformly across contents. An
evolutionary approach implies a radically different model of human intelli-
gence..." (Cosmides et al., 2010, pp. 9007).

Evolutionary reasoning is often thought to lend extra credibility to the cognitive
sciences, by linking psychological hypotheses to physical and biological processes. For
instance, it is the evolutionary perspective on cognition that justifies the search for gen-
erality and universality in psychological processes:

"Perhaps psychological science need not limit itself to the description of em-
pirical regularities observed in the behaviors of the particular, more or less
accidental collection of humans or other animals currently accessible to our
direct study. Possibly we can aspire to a science of mind that, by virtue of
the evolutionary internalization of universal regularities in the world, partakes of
some of the mathematical elegance and generality of theories of that world.
(Shepard, 2001, pp. 601, emphasis added)

Central to the idea that evolutionary reasoning can illuminate the nature of cogni-
tion is the notion of adaptation: some hypothesised feature of cognition is more plausible
if it increases the relative fitness of its possessor. I will refer to this line of reasoning as
an evolutionary rationale for nativism about some psychological domain.

1.3.1 The Evolutionary Defense of Linguistic Nativism

The evolutionary perspective is particularly important to questions of language-related
cognition. It is becoming increasingly clear that the study of language cannot be di-
vorced from the study of its evolutionary history: "[e]volutionary approaches [to lan-
guage], in the broadest sense, are transforming the theoretical terrain" (Evans and Levin-
son, 2009). The general sentiment that the cognitive underpinnings of language are
evolutionary specialisations is widely thought to be well supported:
"...the overwhelming weight of evidence remains that learning of language is an evolved trait unique to humans, that it is an innate cognitive disposition, and hence that it is constrained by its genetic part-causation to learn about a very specific and narrow feature of the world of humans." (Plotkin, 2007, pp. 143)

"It is a fact that grammar is universal in human populations and that our brains have evolved to embody grammar..." (Hinzen, 2014, pp. 97)

The perspective is so widely afforded credibility because, in the case of language, there exists a clear, intuitive, precisely articulated neo-Darwinian explanation for the evolution of its cognitive underpinnings: the hypothesis that language is an adaptation whose cognitive underpinnings evolved under selection pressures relating to communication.

1.3.1.1 The Adaptationist Rationale For Linguistic Nativism

The hypothesis that UG is an adaptation for communication is most closely associated with the work of Steven Pinker, Ray Jackendoff, and Paul Bloom, among others (e.g. Pinker and Bloom, 1990; Pinker, 1994; Jackendoff, 2002; Pinker, 2003; Jackendoff and Pinker, 2005; Culicover and Jackendoff, 2005; Pinker and Jackendoff, 2005, 2009a,b). Even in the earliest of those references, though, these authors generally talk as if they are defending the hypothesis rather than proposing it, so natural and long-standing is the idea. For instance, in their seminal paper Natural language and natural selection, Pinker and Bloom (1990) say of their defence of the hypothesis:

"In one sense our goal is incredibly boring. All we argue is that language is no different from other complex abilities such as echolocation or stereopsis, and that the only way to explain the origin of such abilities is through the theory of natural selection. One might expect our conclusions to be accepted without much comment by all but the most environmentalist of language scientists..."(Pinker and Bloom, 1990, pp. 708)

and in a later paper:

"...that the human language faculty is a complex biological adaptation that evolved by natural selection for communication in a knowledge-using, socially interdependent lifestyle ... might seem to be anyone’s first guess about
the evolutionary status of language, and the default prediction from a Darwinian perspective on human psychological abilities” (Pinker, 2003, pp. 16)

On this account, the evolution of language-specific cognition is straightforwardly comparable to the evolution of other complex structures or behaviours. The force of the hypothesis is strikingly intuitive: i) successful communication affords communicators a fitness advantage; ii) innate language learning machinery that imposes constraints on cognition improves communication by biasing learners toward shared linguistic conventions, therefore; iii) evolution leads to the emergence of strong innate constraints on language. In other words, selection for coordinated communicative convention can explain the existence of strong domain-specific innate cognitive constraints that bias the process of language acquisition. Furthermore, they suggest an evolutionary perspective implies this model of cognition. Though the authors referenced above have been the most vocal proponents of the adaptationist hypothesis, the idea that natural selection for communication has played a role in the formation of language-related cognition is widespread. Pinker and Bloom (1990) cite Bickerton (1981) and Liberman and Mattingly (1989) as fellow adaptationists, and commentaries on their article exemplify claims of support (e.g. Tooby and Cosmides, 1990; Jackendoff, 1990; Newmeyer, 1990; Ridley, 1990; Gopnik, 1990; Sperber, 1990; Broadwell, 1990; Hurford, 1990) and even precedence (e.g. Limber, 1990; Lieberman, 1990). The idea is central to various theories of language (e.g. Nowak and Komarova, 2001; Dunbar, 2003; Corballis, 2003; Hurford, 2011; Scott-Phillips et al., 2011), and the foundation of many formal and computational approaches to the origins of language related cognition (e.g. Kirby and Hurford, 1997; Briscoe, 2000; Nowak and Krakauer, 1999; Nowak et al., 2000, 2001; Mitchener and Nowak, 2003; Baronchelli et al., 2012; Christiansen et al., 2011).

Why is this evolutionary rationale so important to nativist theories of language? Nativist theories explain structure and regularity in language by appeal to innate, domain-specific constraints on cognition. By doing so, the explanation of those properties is being pushed back: ultimately, the value of those theories now depends upon explanations for the existence of whatever cognitive structures are hypothesised to exist. It is worth illustrating this point further. Contrast contemporary instantiations of arch nativist and empiricist approaches to language: Simpler Syntax (Culicover and Jackendoff, 2005) in the nativist camp, and the language as shaped by the brain theory of Christiansen and Chater (2008) in the empiricist camp. Both theories stipulate a wealth of rich innate
cognitive machinery that is more or less isomorphic to the structure of language. Both theories aim to explain how language came to have that structure. The empiricist proposal supposes the cognitive machinery is innate but domain-general. Language has undergone specialisation for cognition: it has been shaped to match pre-existing cognitive constraints by cultural evolution resulting from cultural transmission. The ultimate explanation for that structure lies in whatever processes (unrelated to language) brought about those aspects of cognition\textsuperscript{7}. The nativist proposal supposes the cognitive machinery is innate but domain-specific. Cognition has undergone specialisation for language. As an explanation for the details of language, nativism (but not empiricism), entails commitment to some biological evolutionary process that has brought about cognition that is specialised for language\textsuperscript{8}. The hypothesis that UG arose as an adaptive response to selection pressures relating to communication is the default solution to that explanatory problem. However, the approach is not without its critics. In the next section, I briefly discuss a prominent alternative evolutionary approach, and ask where it leaves the adaptationist rationale for linguistic nativism.

### 1.3.1.2 Discontent with Adaptationist Approaches to Linguistic Nativism

It is important to be clear that nativism does not entail a commitment to adaptationist hypotheses. The two generally go hand in hand, and "[t]he adaptationist position is probably the most widely held view of the origin of UG" (Christiansen and Chater, 2008, pp. 491). However, a plausible adaptationist rationale is evidence in support of nativism, not a necessary component of the thesis. For example, Chomsky famously rejects the idea that UG is an adaptation for communication (e.g. Hauser et al., 2002, 2014; Bolhuis et al., 2014) in favour of evolutionary explanations that appeal to non-adaptationist processes such as exaptation (Gould and Lewontin, 1979), and developmental (see e.g. Spencer et al., 2009; Stotz, 2008) and physical or computational laws (Chomsky, 2005). Chomsky’s earlier scepticism about the adaptationist premise is notoriously vague. However, the rise of minimalism has brought with it more comprehensive formulations of associated evolutionary arguments. Introductions to the non-adaptationist evolutionary

\textsuperscript{7}I use the term ultimate explanation in a non-technical sense, and am not making any claims with respect to the recent heated debates about the status of cultural evolution as an ultimate or proximate cause (e.g. Laland et al., 2011; Dickins and Barton, 2012).

\textsuperscript{8}Note that I am not saying that biological evolution replaces cultural evolution for nativist theories. As I will discuss shortly, the foundation of this thesis will be the idea that cultural transmission is a necessary component of nativist theories in addition to biological evolution.
Chapter 1  Nativism, Evolution, and Culture

theories typical of Chomksyan linguistics and minimalism practitioners, often termed *biolinguistic* theories, can be found in Longa et al. (2012) and Longa (2013), though see for example Chater and Christiansen (2010) or Tallerman (2014) for criticisms. While, as Deacon puts it, "Chomsky’s nonadaptationist view is not, however, widely accepted, even by those who otherwise promote his strong nativist approach to linguistic theory" (Deacon, 2010, pp.9001), the existence of non-adaptationist accounts of language-specific innate machinery highlights two important questions. First: is the adaptationist rationale a coherent explanation for the existence of strong language-specific innate constraints? Second, is the adaptationist rationale the *correct* explanation for the existence of strong language-specific innate cognitive constraints? Of these two questions, the first is by far the most pressing. While answers to the second question may be very difficult to find at present (Lewontin, 1990), answers to the first are in principal obtainable, and determine whether the second question is even worth asking. Much debate between defenders of adaptationist and non-adaptationist theories has concerned the second question. For instance, the exhaustive exchange between Hauser et al. (2002), Pinker and Jackendoff (2005), Fitch et al. (2005), and Jackendoff and Pinker (2005) thoroughly contrasts adaptationist and non-adaptationist accounts of language. However, much of that debate concerns the kinds of language-specific cognitive architecture that need to be explained by an evolutionary perspective, and which would be the *better* explanation for their existence. Consider the following excerpt. It is worth quoting in full since, though it is lengthy, Pinker and Jackendoff’s closing summary is a concise overview of the major disagreements:

"In sum, we find [Hauser, Chomsky, & Fitch’s] case that language is not an adaptation for communication unconvincing. The argument that presupposes the Minimalist Program to conclude that language is too simple to require invoking natural selection is circular, because this is a desideratum that the MP hopes to fulfil (in the teeth of much counterevidence), rather than a discovery it has established. The argument that language is no better designed for communication than hair styles is belied by the enormously greater expressive power of language and the fact that this power is enabled by the grammatical machinery that makes language so unusual. The argument that language is "perfect" or "optimal" has never been stated clearly, and is, by
Chomsky’s own admission, apparently refuted by many “imperfections.” The argument that language is not redundant is false in every domain in which it can be evaluated. Finally, the suggestion that the recursive power of language arose as a simple co-opting of recursion in other cognitive systems such as navigation or number encounters numerous problems: that navigation is not discretely infinite; that recursive number cognition is parasitic on language rather than vice-versa; and that language maps among recursive systems rather than being a straightforward externalization of a single recursive system. The alternative in which language is an adaptation for the communication of knowledge and intentions faces none of these problems.” (Pinker and Jackendoff, 2005, pp. 231)

Of the arguments rehearsed in that quote, none concerns the premise, or the viability, of the adaptationist evolutionary rationale for an hypothesised UG; none questions whether biological evolution through natural selection for improved communication would lead to strong innate constraints. Rather, the majority concern the nature of the to-be-explained capacity: whether or not language is redundant or optimal, whether or not language is well designed for communication, whether UG contains rich specifics or just recursion, etc. The competing views of the evolutionary process behind language largely reflect different views about what language is: neither party’s evolutionary model suits the other’s language model. This line of argument against adaptationist reasoning is roughly as follows: the kind of UG that adaptationism predicts is not the kind of UG we are trying to explain. For minimalists, adaptationist evolutionary explanations are unnecessary for their explanadum: the core computational procedure required to represent hierarchical structure, merge. It is highly instructive to note that, despite hypothesising alternative evolutionary explanations for merge, the adaptationist premise is nevertheless broadly accepted as viable (for broader aspects of language-related cognition other than merge) even among this school:

"...we take for granted that the large set of complex mechanisms entering into FLB are adaptive in some broad sense, having been shaped by natural selection for, among other things, communication with other humans. We find this idea...[un]controversial...” (Fitch et al., 2005, pp. 189)

"...numerous areas of agreement...remain. Agreed is the argument that FLB, as a whole, evolved and functions as a human-specific adaptation with several areas of current utility, one of which is clearly communication with con-
specifics.” (Fitch et al., 2005, pp. 205)\footnote{In these excerpts, FLB refers to the language faculty in a broad sense. That is, all aspects of cognition that contribute to language, rather than just the unique computational capacity for building recursive syntactic structures.}

In addition to the Chomskyan eschewing of adaptationism, there have been more general arguments to the effect that adaptationist accounts predict the wrong kind of UG. For instance, Christiansen and Chater (2008) articulate what I will term the *problem of dispersion*. The argument is as follows. If natural selection for communication builds strong innate domain-specific constraints, the fact that human populations have been dispersed around the globe for thousands of years predicts that distinct populations will possess distinct UGs, each adapted to the local environment. However, it is a general point of consensus, and a central assumption of the nativist endeavour, that UG is uniform among the species. The same premise (that adaptation leads to local specialisation) leads to another related problem, which I will refer to as the *problem of plasticity*. Namely, if adaptation among isolated groups leads to independent specialisation, we would not expect any child to be able to learn any of the world’s languages with equal ease: localised biases would favour certain languages over others. Yet it is clear that this is not the case, since children appear to be able to learn any language with equal ease, regardless of where they were born (see e.g. Fitch et al., 2005). These problems represent different faces of a pretty substantial apparent paradox that runs deep in the study of language, and receives little resolution in the adaptationist defence of UG: how can we square universality in cognition with variation in culture? A satisfying evolutionary perspective on language should ideally offer some resolution to conflicts like this.

### 1.3.1.3 The Viability of the Proposal

In addition to the arguments discussed in the previous section, there have been some more fundamental concerns about the *viability* of the adaptationist evolutionary rationale for nativism: *does the adaptationist account actually predict strong domain-specific constraints?* For instance: the hammer and nail problem (see Papineau, 2006) states that multi-faceted traits are difficult to evolve because no sub-component has value in the absence of the remaining components. Pinker and Bloom (1990) address other similar concerns. For example, one potential problem relates to the value of incrementally
stronger constraints: if, as with other complex traits, improvements over evolutionary time occur in small gradual steps, are the intermediate stages useful enough to be selected for?; how valuable is 5% of an innate constraint? Another similar problem concerns the value of cognitive innovations that are not \textit{shared}: when innate constraints first emerge, would they improve the possessor’s ability to communicate with others who aren’t subject to the same constraints? These are interesting points that I will explore further during later chapters. However, in the next section, I turn to a more fundamental criticism of the adaptationist evolutionary rationale for linguistic nativism. Adaptationist hypotheses, and the nativist theories they support, have come under new fire relating to a topic that is rapidly gaining attention in the language sciences: cultural evolution.

\subsection*{1.3.2 Cultural Evolution}

Recent years have witnessed an explosion of interest in approaches to the origins of language (e.g. Kirby et al., 2014; Steels, 2011; Fay et al., 2010; Christiansen and Chater, 2008), and of human behaviour more generally (e.g. Richerson and Boyd, 2005; Mesoudi, 2011), that take seriously the role of cultural transmission and evolution. Language is culturally transmitted through an iterated cycle of expression, inference, and induction. This transmission mechanism induces an evolutionary process with interesting dynamics of its own: languages evolve culturally, simply by deed of being repeatedly passed from one learner to another. Despite being an uncontroversial fact about the nature of language, this important property has been almost entirely absent from traditional nativist linguistic theories. This may well turn out to have been an important omission. Virtually all contemporary theories of language, and nativist theories in particular, propose to explain aspects of the population-level structure of language, at least in part, by appeal to properties of the cognitive architecture of individuals. It is incontestably true that it is cultural transmission that links these two phenomena. At the very least, cultural transmission is the \textit{mechanism} through which properties of individual cognition can become expressed in language (Kirby, 1999), and therefore must have a place in any theory that relates universals of language and properties of cognition. More troubling to traditional nativist theories is the possibility that the nature of cultural transmission actually influences the resultant structure of language. Rather than simply providing means for the direct expression of cognitive constraints on the legal structure of lan-
guage, the cultural process may itself impose constraints on linguistic form. If this is the case, and the relation between cognition and linguistic structure is made non-trivial by the intervening cultural process, as has been claimed (e.g. Kirby et al., 2004), the consequences for nativist theories of language may be profound. First, the inference from linguistic universals to isomorphic cognitive constraints would be compromised; the argument from universal tendencies introduced above would be defused (see e.g. Smith and Kirby, 2008). More pertinently, though, the adaptationist evolutionary rationale for linguistic nativism would be severely compromised (Smith, 2011; Smith and Kirby, 2008; Kirby et al., 2007).

The general idea that cultural processes have influenced human evolution has been taken seriously for at least several decades now (e.g. Waddington, 1961; Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985; Durham, 1992; Laland et al., 2010). Several authors have suggested that the interaction between culture and biology makes the evolution of language one of the richest remaining problems in contemporary science (e.g. Christiansen and Kirby, 2003; de Boer, 2014). Nevertheless, cultural evolution has received little attention among those who propose nativist theories for language. These fundamental concerns will be central in this thesis, and I will return to their discussion shortly. Before doing so, however, it will be useful to consider an influential critique of the adaptationist hypothesis for language that results directly from a perspective that accommodates cultural evolution.

### 1.3.2.1 Culture and Linguistic Nativism

Christiansen and Chater (2008) (C&C) present several arguments against linguistic nativism and associated evolutionary rationales that relate directly to the consequences of cultural transmission and evolution. Most are of the form outlined in section 1.3.1.2: they question the evidence for strong domain-specific innate constraints of the kind predicted by adaptationist (and non-adaptationist) approaches, and argue for an empiricist model of the innate foundations for language. However, in addition to these suggestions, C&C also formulate a critique of the viability of the adaptationist rationale for nativism: they argue that the evolutionary process envisaged by Pinker and Bloom would not lead to the formation of rich domain-specific innate constraints on learning. Their main proposal is known as the moving target argument, and is based on the premise that language change proceeds at a rate much greater than biological
evolution. Linguistic conventions, they claim, are a moving target that biology cannot keep up with: the speed of cultural evolution, relative to that of biological evolution, prohibits cognitive adaptation to linguistic convention\textsuperscript{10}. On the basis of this argument, alongside others, C&C claim that nativist theories have no credible evolutionary rationale. The argument has received considerable attention, and is widely thought to be a serious complication for adaptationist accounts of language (e.g. Evans and Levinson, 2009; de Boer, 2014; Számadó and Szathmáry, 2012, - also see commentaries on C&C’s original BBS article). Furthermore, computational simulations (e.g. Chater et al., 2009; Baronchelli et al., 2012) and mathematical models (Baronchelli et al., 2013) of the basic premise (and several variations) bear out the moving target idea: fluctuating environments compromise adaptation (also see Boyd and Richerson, 1985)\textsuperscript{11}. Like the adaptationist rationale for linguistic nativism, the moving target argument is strikingly intuitive: if languages themselves evolve, how can we have been wired to learn them?

The case that cultural evolution has profound implications for the viability of the evolutionary rationale underpinning linguistic nativism is indeed compelling: the traditional synthesis of neo-Darwinian reasoning and linguistic nativism must surely be reconsidered in light of the cultural turn. However, the conclusion that C&C are forced to draw - that culture simply shuts down the biological evolutionary adaptive process - may only apply to a limited class of cognitive tasks. In fact, I will argue in this thesis that the evolutionary process induced by biological adaptation to a culturally evolving trait may be one of the richest topics to fall out of the study of cultural evolution. As others have done (e.g. Pagel and Atkinson, 2008; Fitch, 2008; Barrett et al., 2008; Számadó and Szathmáry, 2012), I will argue that the viability of cognitive adaptations for language may be dependent in complex ways upon the dynamics of gene-culture co-evolution.

1.3.2.2 Gene-culture Co-evolution and Language

Gene-culture co-evolutionary processes are known (see e.g. Boyd and Richerson, 1985; Laland and Brown, 2011) to often exhibit counter-intuitive properties: while C&C may be right to rule out cognitive adaptation to certain kinds of cultural behaviour, there may be others that follow richer evolutionary dynamics because of culture. Though it

\textsuperscript{10}I focus on C&C’s version of this argument because it is contemporary, and has given rise to considerable debate in the field. It is worth noting, however, that Deacon (1997) articulated essentially the same argument, and Briscoe (2000) derived a co-evolutionary model (discussed in chapter two) to examine its premise, almost a decade earlier.

\textsuperscript{11}These models, and others like them, will be the subject of chapter two.
is clear that the cultural evolution of language may have profound implications for the biological evolution of cognitive substrates (e.g. Smith and Kirby, 2008), it is far from clear that those implications necessarily rule out cognitive adaptation. Culture does more than alter the outcome of the evolutionary process: it fundamentally changes the nature of the process:

"...when evolution results in learners able to transmit information directly between themselves - ‘short-circuiting’, as it were, the genetic channel of information transmission...[t]his alters evolution in a dramatic and significant manner; it forms a major transition in the evolution of life on earth...and it adds further complexity to the causal structures of behaviour in those species able to learn from one another." (Plotkin, 2007, Ch. 6, pp. 229)

Our evolutionary predictions about the cognitive underpinnings of language must account for this complexity. Human culture is "...awesomely powerful, sufficiently so to have determined human evolution" (Plotkin, 2007, pp. 236). Christiansen and Chater’s arguments for the importance of culture in biological evolution are the tip of an iceberg: that is, C&C may actually understate the magnitude of their observation (Smith et al., 2008b), for several reasons.

For example, C&C are explicit about the fact that their argument against the viability of the adaptationist evolutionary rationale for nativism only applies to linguistic conventions that vary as a result of cultural or historical factors and are not significantly influenced by the cognitive biases of language learners (see e.g. Chater et al., 2009). This is crucial, for at least two reasons. First: while there may well be aspects of language whose properties are largely caused by factors other than learners’ cognitive constraints, those are not generally the explanandum of nativist theories. As is clear from the preceding sections of this chapter, it is a central premise of nativist linguistic theories that the structural properties of language are brought about by cognition. Second: the dynamics of gene-culture co-evolution are richest, and have the greatest potential to deliver surprising results, where there is reciprocal causation between biology and behaviour (Laland et al., 2011). In this case, that corresponds to the idea that, in addition to cognition evolving in response to pressures brought about by language, language also evolves in response to cognition. C&C’s moving target critique does not apply to an evolutionary process that includes reciprocal causation: consequently, they are
forced to accept the viability\textsuperscript{12} of the adaptationist evolutionary rationale in precisely the case that is most important to nativist theories, and most likely to be subject to complex and counter-intuitive dynamics. In fact, this issue is the foundation of what has perhaps been the most influential critique of adaptationist reasoning in general (e.g. Lewontin, 1983). Lewontin’s famous rejection of standard adaptationism, which has been a considerable influence on Chomsky’s anti-adaptationist position regarding UG, revolves around the fact that traditional adaptationist theorising ignores reciprocal causation (see Ch. 4 of Plotkin, 2007). This thesis will be built on the alternative premise that complex evolutionary dynamics, such as those implied by reciprocal causation, present a rich but tractable object of study, rather than an \textit{a priori} roadblock to evolutionary insights. That cognitive evolution in a cultural species involves reciprocal causation should encourage a re-formulation of our evolutionary predictions in the light of culture, rather than a rejection of evolutionary reasoning about cognition \textit{per se}.

Reciprocal causation is not the only factor that potentially complicates cognitive adaptation to a cultural behaviour. Another concerns the nature of the problem to which the biological evolution of innate constraints is a solution. In their defence of the adaptationist evolutionary rationale for nativism, Pinker and Bloom (1990) make much of Liberman and Mattingly’s (1989) observation that language, like any other communication system, relies on \textit{parity}: to be communicatively useful, everybody must speak the \textit{same} language. Pinker and Bloom primarily use this fact as a defence against the accusation (see e.g. Piattelli-Palmarini, 1989) that functionally arbitrary linguistic conventions, which are central to linguistic nativism, are unlikely adaptations: the communicative nature of language imposes the requirement for parity, which can be well satisfied by functionally arbitrary structures, so long as they are shared.

Chater and Christiansen (2010) generalise Liberman and Mattingly’s observation, and claim that problems of induction can be divided into two kinds: natural problems of \textit{n-induction}, in which the goal is to discover a naturally occurring truth that is independent of the learners trying to discover it, such as physical law, and; cultural problems of \textit{c-induction}, in which the goal is simply to arrive at knowledge that is co-ordinated with other learners. Language is a paradigm example of \textit{c-induction}. Chater and Christiansen primarily use this observation to question the APS. They reason that

\textsuperscript{12}If not the merit, which is questioned on other grounds. See simulation 3 in Chater et al. (2009), discussed at length in the next chapter, for the details of why adaptation is not compromised in the same way under this assumption.
language acquisition has generally been studied as if it were a problem of n-induction: if it were an n-induction problem, and the goal of acquisition were induction of a learner-independent body of knowledge, any "incorrect" inferences (that might result from impoverished data, for example) would be problematic for the learner. Acknowledging that it is a problem of c-induction can, they argue, diffuse nativist claims: the goal of acquisition is only shared knowledge, and learners' errors work in their favour, because other learners equipped with similar cognitive resources are likely to make similar "incorrect" inferences, therefore we need not hypothesise specialised cognitive resources that overcome these errors. More important to my concerns, they also suggest that problems of c- and n-induction imply different evolutionary processes. That adaptation to problems of n-induction can lead to rich, strongly constraining cognitive architectures is a natural prediction:

"When formalized at a sufficient level of abstraction, mental principles that have evolved as adaptations to principles that have long held throughout the universe might be found to partake of some of the generality of those prior principles...perhaps even attaining the kind of universality, invariance, and formal elegance (if not the quantitative precision) previously accorded only to the laws of physics and mathematics." (Shepard, 2001, pp. 581-582)

The notion of rich cognitive adaptations that are abstract, universal, invariant, and formally elegant is strongly reminiscent of classic claims about UG. But Shepard’s mental principles belong to a different class of evolutionary problem. That language is a problem of c-induction is another illustration that its cultural nature implies an evolutionary model (for its cognitive foundations) more intricate than existing proposals tend to recognise. Selection for adaptations that solve a problem of c-induction is frequency dependent: the communicative value of any evolved cognitive specialisation depends upon its frequency in the population. Like reciprocal causation, frequency dependent selection is known to induce complex non-linear dynamics in evolutionary systems (see e.g. Nowak, 2006).

Thus, there are fundamental aspects of language that result from its cultural nature which must be worked into an evolutionary account of the origins of language-specific cognition, but have hitherto been under-explored:

"In my opinion the greatest failure in this enterprise [explaining the origins of language] has been a failure to appreciate the full complexity of the
evolutionary problem it poses. This is reflected in the remarkably simple evolutionary logic typically offered by standard accounts of language competence in humans...we tend to conceive of language in a way that ignores the complex self-organising and evolutionary dynamics that form the very essence of its design logic.” (Deacon, 2007, pp. 81)

The evolutionary rationale behind linguistic nativism is a hypothesis about a gene-culture co-evolutionary process with some specific properties. What are the prospects for understanding how culture influences cognitive evolution, and the consequences for nativism? Recently, there has been a surge of interest in this topic. There is huge interest in integrating biological and cultural approaches to language (see e.g. Mesoudi et al., 2011, and papers in that special issue), and in understanding the consequences of cultural transmission for the evolution of human behaviour more generally (see e.g. Heyes, 2012; Smith et al., 2008a, and papers in those special issues). Levinson (2006) articulates a common sentiment concerning the relationship between nativism and cultural evolution:

"What is right about simple nativism is that it insists on the prestructuring of our mental abilities. What is wrong about it is that it minimizes or ignores the role of ontogeny and learning, and minimizes the very stuff of our evolutionary success, namely, the cultural variation that is our special system for rapid adaptation to differing environments.” (Levinson, 2006, pp. 14)

There is a growing consensus that learning, culture, and biology are inextricably linked in the origins of human behaviour (Plotkin, 2007), and that co-evolutionary dynamics defined by the interaction of these processes underpin our unique cognitive capacities: "a co-evolutionary spiral is not just feasible but in our opinion it is the likeliest explanation behind the current complexity of human brain, language, and culture” (Számadó and Szathmáry, 2012, pp. 165). This trend, alongside the contemporary perspective that strong nativism is not the only possible view of how domain-specific cognition can support language (recall the ontology of possible cognitive theories discussed in section 1.2), suggests that our understanding of how evolution can shape cognition should be re-evaluated. The co-evolutionary model underlying the evolutionary rationale for nativism should be re-formulated in light of these contemporary perspectives. In the final section of this chapter, I argue that a recently developed formal framework can allow us to do just this. Earlier, I suggested that Bayesian computational
models of cognition can capture the contemporary approach to innateness in cognition through arbitrary prior distributions. By combining principled statistical inference and prior knowledge, Bayesian models offer a well understood formalism for representing the link between biology and learning. Crucially, recent developments mean that Bayesian models now also offer formal representations of culture.

1.3.2.3 Bayesian Iterated Learning

Models of Bayesian iterated learning (e.g. Griffiths and Kalish, 2007; Kirby et al., 2007; Smith, 2009; Burkett and Griffiths, 2010) formalise the dynamics of cultural transmission induced when Bayesian learners learn from each other. Iterated learning is one kind of cultural transmission, whereby individuals learn a behaviour by observing that behaviour in another individual who learned in the same way (Kirby et al., 2014). Bayesian models of iterated learning begin with a well understood model of cognition, combine it with a well understood model of cultural evolution, and derive the dynamics that result from long term transmission among learners of this kind. The iterated learning paradigm (see e.g. Kirby et al., 2014, for a review) has a history of application to questions concerning the emergence of linguistic structure. At its core is the idea that the process of cultural transmission may influence the structure of language. There are at least two related hypotheses about the way this can happen. First, the transmission process itself might inherently favour certain linguistic structures over others: recursive syntax (e.g. Kirby, 2002), morpho-syntactic regularity (e.g. Kirby, 2001), compositionality (e.g. Smith et al., 2003), or, more generally, compressible representations (e.g. Tamariz and Kirby, 2014) may be natural outcomes of iterated learning, for instance. In line with this hypothesis, cultural evolution is often (cogently) treated as an alternative to nativist theories of language (see e.g. Evans and Levinson, 2009). Second, the cognitive biases of individual learners might, through cultural transmission, become amplified, attenuated, or otherwise transformed. My focus here is squarely on the second of these hypotheses. In this thesis, I will treat cultural evolution as a necessary component of nativist theories, rather than an alternative.

Perhaps the most important contribution of Bayesian models of iterated learning has been to relate long-term, population-level properties of language to the inductive biases of individual learners; to explicitly relate the outcomes of culture to the constraints imposed by biology. Griffiths and Kalish (2007) find two general patterns for
this relationship, and show how minor differences in the individual learning process can switch between these patterns. First, in some circumstances, cultural transmission can lead to long term population-level distributions that faithfully reflect the cognitive biases of individual learners. Second, in other circumstances, learners’ cognitive biases can become amplified over the course of cultural transmission: the long term population-level distribution of linguistic forms can be more strongly skewed toward favoured variants than would be predictable from learners’ innate biases. In this case, culture can fill the explanatory gap inherent to theories of minimal nativism described in section 1.2.1. Of the two, the second, which I will refer to as the amplification effect, seems most likely to complicate the evolution of the language faculty (Kirby et al., 2007; Smith and Kirby, 2008). The amplification effect breaks the direct link between learners’ innate biases and the linguistic forms they end up acquiring. Since natural selection acts over the latter but influences the former, the breaking of this link implies potentially non-trivial evolutionary dynamics. Nevertheless, both possible patterns for the relationship between cognition and culture may have consequences for the evolution of language-specific innate cognitive biases.

Nativist theories of language, and the evolutionary models upon which they depend, have ignored culture. I have argued in this section that the basic evolutionary rationale for linguistic nativism does not account for several fundamental properties of language that follow from cultural transmission. These properties imply complex co-evolutionary dynamics. The basic adaptationist proposition should, in the light of culture, be reformulated roughly as follows: the dynamics of gene-culture co-evolution, in a problem of c-induction, subject to reciprocal causation, under natural selection for communication, favor the biological evolution of strong innate constraints on cognition. Though many formal models of gene-culture co-evolution exist, none, I will argue in the next chapter, comprehensively tackles this particular scenario. The logical viability of the adaptationist evolutionary rationale for linguistic nativism therefore remains untested.

1.4 Summary

This concludes the chapter. In summary: nativism is a long-standing philosophical doctrine that runs deep in cognitive sciences. It has profound implications for our understanding of human nature, and is closely tied to evolutionary questions. Linguistic nativism is thought to be the gold standard for nativist theories: it leans on a num-
ber of persuasive arguments, including a well-rehearsed evolutionary rationale, and has enjoyed unparalleled success and influence. Language has been used as a flagship example that evolution can build domain-specific rich innate structures for complex arbitrary behaviours.

Many have questioned linguistic nativism on a range of fronts (Scholz and Pullum, 2002). Some have argued against the APS, claiming that the argument is incoherent (e.g. Cowie, 1999), or that the stimulus is not impoverished (e.g. Pullum and Scholz, 2002), or that domain-general learning mechanisms are a sufficient solution (e.g. Perfors et al., 2011b), for instance. Others have argued against the AUT, on the basis that true linguistic universals don’t exist (e.g. Evans and Levinson, 2009), or that linguistic universals reflect functional pressures rather than cognitive constraints (e.g. Newmeyer, 2007). Likewise, the adaptationist rationale for linguistic nativism has been challenged in favour of alternative evolutionary mechanisms, such as non-adaptationist processes (e.g. Hauser et al., 2002), or cultural evolution (Christiansen and Chater, 2008). In this thesis, I take a different approach: I’ll address the evolutionary rationale for (and thus the credibility of) linguistic nativism on its own terms. I’ll accept the basic nativist model for language, and the adaptationist premise for the formation of innate knowledge, but make one seemingly minor addition: into both of these frameworks, I introduce cultural transmission. I have argued that this approach implies a specific evolutionary model, based on gene-culture co-evolutionary interactions, that remains understudied. The central goal of this thesis is the construction and analysis of mathematical models of this process, in order to ask how culture shapes the evolution of cognition, and what this means for nativist theories.
Chapter 2

Co-evolution: A Review of Existing Models

In the previous chapter, I questioned the credibility of a flagship case for the marriage of nativism and evolutionary reasoning in contemporary psychological science: the adaptationist evolutionary rationale for linguistic nativism. I did so on the grounds that language, like many human behaviours, is culturally transmitted: as such, whether or not evolution can build innate language-specific cognitive constraints depends upon co-evolutionary dynamics with specific properties, such as reciprocal causation and frequency dependent selection. I declared my goal in this thesis to be the construction and analysis of formal models that can illuminate these dynamics in order to ask: how can culture shape the evolution of cognition?

Many co-evolutionary models relevant to this question already exist. In what follows, I review that literature in detail. Ultimately, I will conclude that none comprehensively explores the co-evolutionary process underlying the evolutionary rationale for linguistic nativism. The review is split into two relatively self-contained parts. Section 2.1 reviews the most relevant models and findings from the gene-culture co-evolution literature. Section 2.2 details co-evolutionary models specifically designed to resolve questions concerning the evolution of language-related cognition.
2.1 Gene-Culture Co-evolution

The interaction between genetic and cultural inheritance in evolutionary systems is now widely acknowledged to have played a crucial role in human evolution (Richerson and Boyd, 2005; Laland et al., 2010), and perhaps even been the dominant mode of evolution in the history of our species (Laland, 2008). Gene-culture co-evolution is now thought to be so pervasive that some are even calling for relevant revisions to textbook evolutionary theory (e.g. Laland et al., 2011; Mesoudi et al., 2013). There exists much ground-breaking empirical study of gene-culture co-evolution: specific phenomena whose co-evolutionary origins have been studied empirically include human handedness (Laland et al., 1995), adult human lactase persistence (e.g. Bersaglieri et al., 2004), birdsong acquisition and evolution (e.g. Kagawa et al., 2014), alcohol consumption in humans (e.g. Chen et al., 1999), human morality (e.g. Mrazek et al., 2013), and even tool use in dolphins (Kopps et al., 2014)\(^1\). However, the core of this research enterprise is mathematical modelling: since the 1980s (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1985), our understanding of co-evolutionary dynamics has largely grown out of a series of related formal analyses generally known as models of gene-culture co-evolution (Laland and Brown, 2011). Though the fingerprints of this process continue to be discovered in new domains (Richerson et al., 2010), classic models tend to concern the origins of traits such as social learning (e.g. Boyd and Richerson, 1985; Perreault et al., 2012), human adult lactose tolerance (Holden and Mace, 1997), altruism and punishment (Gintis, 2003; Boyd et al., 2003), incest taboos (Aoki and Feldman, 1997), birdsong (Lachlan and Feldman, 2003), and sex ratios (Kumm et al., 1994). Some of these models bear more relevance than others to the origins of cognitive biases for culturally transmitted language. Here I review details of the more relevant topics and the models that underpin them, beginning with the case that has received most attention, social learning (section 2.1.1), then turning to birdsong (section 2.1.2), direct bias (section 2.1.3), and sign-languages (section 2.1.4). Each of these cases has been studied with models of co-evolutionary processes that resemble, but are crucially distinct from, the co-evolutionary process behind the evolutionary rationale for linguistic nativism.

\(^1\)For a recent review of gene-culture co-evolution and its applications, see Laland et al. (2010).
2.1.1 Social Learning

Broadly speaking, social learning represents "the transfer of learned information from one individual to another" (Kendal et al., 2009, pp. 210). It is endemic to human activities (Richerson and Boyd, 2005), and widespread among other animals (Galef and Laland, 2005). The origins of social learning are the subject of a rich body of co-evolutionary models dedicated to the question: how can co-evolution equip us with cognitive equipment specialised for learning from and among other minds? These questions are hotly debated, since social learning is thought be the foundation of much human success (Richerson and Boyd, 2005), and to even represent a major transition in the evolution of life (Plotkin, 2007). Interestingly, debates concerning the evolution of the cognitive basis of social learning roughly parallel those concerning the evolution of innate linguistic knowledge.

Historically, a predisposition to learn socially was thought to be a straightforwardly plausible adaptation, since it provides a mechanism for acquiring useful information without facing any costs that might be associated with uncovering that information individually (Kendal et al., 2009). Co-evolutionary considerations have shown that this intuition may miss some of the complexity introduced by variable environments. In brief, populations full of social learners (information scroungers), and lacking in individual learners (information producers), can quickly find themselves holding on to outdated knowledge. Environmental variation induced by cultural evolution is precisely what Christiansen and Chater (2008) claim, in their moving target argument, also prohibits the biological evolution of innate linguistic predispositions, despite the intuitive plausibility of those adaptations.

However, thanks to a series of co-evolutionary analyses, our understanding of how environmental variation prohibits the evolution of biases for social learning is fairly comprehensive: fluctuating environments compromise the utility (Whitehead and Richerson, 2009; Kameda and Nakanishi, 2002), and therefore the evolution (Feldman et al., 1996), of social learning; environmental change can lead to population collapse in realistically variable scenarios (e.g. Whitehead and Richerson, 2009); and it necessitates that some proportion of the population be information producing individual learners, otherwise, Roger's paradox states that a bias to learn socially will afford no fitness benefit over individual learning (Rogers, 1988). Despite these complications, co-evolutionary models have shown that social learning biases can evolve. These models generally fall...
into two categories: those that search for conditions under which the basic variety of social learning (i.e. copy others at random) might evolve (section 2.1.1.1), and those that consider enriched versions of social learning (section 2.1.1.2), such that an individual’s learning behaviour is guided by rules, or social learning *strategies* (Laland, 2004).

### 2.1.1.1 Vanilla Social Learning

There exists a lively research program concerned with revealing conditions under which co-evolution can lead to the fixation of innate biases to learn socially rather than individually. Here I focus on two co-evolutionary models: one derived by Boyd and Richerson (1985) in their magnum opus *Culture and the Evolutionary Process*, and one contemporary model which re-casts the problem in inferential terms.

In chapter four of *Culture and the Evolutionary Process*, Boyd and Richerson (1985) explore the biological evolution of a propensity to learn socially rather than individually. In this model, learners must acquire behaviour on the basis of two sources of information: behaviour observed in cultural models, and information gleaned individually from the environment. Induction of behaviour that is consistent with the state of the environment is favoured by selection. Boyd and Richerson describe the to-be-learned behaviour as a *quantitative trait*: the representation a learner induces can be characterised by a single value in some continuous range. Likewise, the state of the environment is a single value from the same continuous range of possibilities. Social learning involves observing the behavioural values exhibited by *n* social models form the previous generation. Individual learning is characterised by what Boyd and Richerson call *guided variation*. In intuitive terms, guided variation assumes that some aspect of the individual learning process allows the learner to rationally mould her behaviour such that it better suits the current environment, and pass that behaviour on. The relative contributions of individual and social learning to the learner’s ultimate behaviour are weighted by a genetically inherited trait: this is the cognitive bias whose evolution is being modelled. Two significant analyses are relevant here, corresponding to cases of environmental stability and variability.

Assuming a stable environment and strong guided variation (individual learning finds the better behaviours quickly and reliably), in the long run selection generally favours populations who prioritise social learning over individual learning: because cultural evolution by guided variation is fast, environment-appropriate behaviours
emerge quickly, which may then be copied by individuals who are disposed to learn socially, and genetic evolution slowly fixes this propensity in the population.

Two kinds of variable environment are considered: *heterogeneous*, in which subsets of the population at any generation experience different environmental states, and *fluctuating*, in which all individuals at a given generation experience the same environmental state, but this state varies between generations. Results of these analyses do not differ qualitatively, so summarising only the second kind is sufficient. In short, a bias towards social learning in varying environments is the outcome of gene-culture co-evolution where two conditions hold: 1) individual, asocial learning is costly or inaccurate, and 2) environmental variability is modest. In sum, Boyd and Richerson’s models collectively show that gene-culture co-evolution can lead to nativisation of a bias for social learning in the face of cultural evolution, but only under quite specific assumptions about how culture evolves and how costly are the respective biases. As Christiansen and Chater suggest, rapid cultural change is an obstacle to cognitive adaptation.

More recently, Perreault et al. (2012) present a co-evolutionary model that is notable for several reasons. Theirs is a Bayesian model of social learning, in which learners must adopt one of two behaviours. In response to the common criticism that social and individual learning are in reality not separate mechanisms, but have been widely modelled as such, Perreault et al. assume a single type of learner characterised by a single inferential process, which simultaneously has access to information gleaned socially and individually. The learner must balance the relative contributions of these evidential sources during inference. The details of this balance are specified by a genetic component which is allowed to evolve by natural selection. As in Boyd and Richerson’s models, learners gain fitness benefits by adopting a behaviour which is consistent with the current environmental state, which across generations fluctuates back and forth between two alternatives.

Two main findings emerge from the analysis. Firstly, standard results from previous models (e.g. Boyd and Richerson, 1985) are recovered: inaccurate environmental information and slowly changing environments favour the evolution of a bias for social learning. Secondly, where learners observe more than one social model alongside their asocial observations, a conformity-bias effect is observed\(^2\). This model is partic-

---

\(^2\)Note that this result parallels an earlier finding reported by Smith (2009), in the neighbouring Bayesian iterated learning framework, which I will discuss during later sections of the thesis.
ularly relevant because it concentrates on the long-term co-evolutionary consequences of aspects of an inferential procedure: it shows that even where evolution is able to tinker with the learning process directly, a variable environment can prevent cognitive adaptation. Furthermore, it shows how at least one aspect of the cultural transmission process (the number of observed teachers) can have significant co-evolutionary consequences.

2.1.1.2 Social Learning Strategies

Many co-evolutionary models exist to explore the emergence of biases that guide enriched social learning in order to navigate the potential traps that can damage populations of indiscriminate social learners. Laland (2004) refers to these rules as social learning strategies, which he partitions into "when-strategies" and "who-strategies". "When-strategies" generally encourage social learning in advantageous conditions: "copy when asocial learning is costly", and "copy when uncertain" are two such strategies described by Laland. "Who-strategies" tend to dispose the learner to engage in social learning if certain conditions about other individuals are met: "copy successful individuals" and "copy kin" are typical examples (Laland, 2004). Many such strategies and their evolutionary stability have been, and continue to be, explored in co-evolutionary models. A recent review, and comparative co-evolutionary analyses, of several common strategies can be found in Kendal et al. (2009). In short, numerous simple strategies that enrich social learning capacities offer evolutionary benefits, and prevent the evolutionary dead-ends sometimes associated with vanilla social learning, such as information stagnation or population collapse. Such strategies can also co-evolve with social learning.

The set of strategies explored, and associated co-evolutionary literature, is too large to review in any principled way here, given the limitations restricting its relevance for nativist theories of language, which I will outline shortly. In brief, the intuition that copying is a valuable behaviour can be upheld in co-evolutionary terms when simple inferential biases are available to evolution. Again, these models demonstrate repeatedly that minor differences in the inferential process can lead to dramatic differences in the long term behaviour of the co-evolutionary system. This observation has important

---

3 It is also particularly relevant because it is one of very few gene-culture co-evolutionary models that assume Bayesian learning.

4 The overwhelming depth and diversity of these models in fact recently encouraged an unusually comparative co-evolutionary analysis in a common framework: a public tournament! Insights from this competition can be found in Rendell et al. (2010), Rendell et al. (2011), and Fogarty et al. (2012).
consequences for making predictions about the evolution of innate linguistic knowledge. Language acquisition is characterised by several foundational properties, such as the fact that inference acts over data from other learners, and the ubiquity of inferential ambiguity thanks to sparse observations. Models of the origins of social learning strategies warn that these properties should be accounted for when making co-evolutionary predictions about language.

2.1.1.3 Social Learning Models: A Brief Discussion

This concludes the review of co-evolutionary models of the origins of social learning. While fascinating in their own right, and vital reference points when thinking about co-evolutionary dynamics, findings from this program cast but a limited light on the evolutionary credibility of linguistic nativism, for two major reasons. First, in these models the target of learning, whether variable or fixed, is generally independent of the learners who reason about it; environmental states and their fluctuations, for instance, are generally treated as regularities brought about by some factor external to the learner. In Chater and Christiansen’s (2010) terms, these models study cognitive evolution in problems of n-induction.

Second, the cognitive adaptations explored in these models belong to a class of inferential biases that are, crucially, distinct from those ordinarily assumed by nativist philosophies. Biases concerning whether or not to copy, who to copy, when to copy, etc., do not act to substantively constrain learners’ prior expectations about individual representations, as language-specific innate constraints do. Learners in these models do not generally a priori prefer one particular environmental state over another; their preferences concern which among several sources of evidence deserve attention. For these reasons, gene-culture co-evolutionary models of social learning do not describe the process that underpins the evolution of language-specific innate constraints.

Overall, the major insight from social learning models is: where systems of knowledge are transmitted culturally, properties of an individual’s cognitive architecture are not the only factors that can influence behaviour. These models show that cultural transmission casts doubt on the inference from universals of behaviour to isomorphic cognitive universals. Recall the following passage from Locke quoted in chapter 1:

"This argument, drawn from universal consent, has this misfortune in it,
that if it were true in matter of fact, that there were certain truths wherein all mankind agreed, it would not prove them innate, if there can be any other way shown, how men may come to that universal agreement in the things they do consent in; which I presume can be done.” (Locke, 1689, p. 12)

Locke presumed right: we now know that social learning, or cultural transmission, is one such “other way”. If the nativist could find a population of social learners universally endowed with some behaviour of interest and look inside their heads, he would, according to the models just reviewed, likely be surprised to find little innate representation of that behaviour, only a predisposition to copy others. Only the dynamics of co-evolution, which connect multiple evolutionary streams over multiple time-scales, could explain such a scenario. In the following sections, I consider several related models of gene-culture co-evolution which directly explore the evolution of substantive prior knowledge.

2.1.2 Learned Birdsong

Learned birdsong is fast becoming a common topic of inquiry for those concerned with the origins of language (e.g. Berwick et al., 2011; Bolhuis and Everaert, 2013). The reasons for this are clear: it is uncontroversially culturally transmitted, through vocal imitation and auditory learning; it often exhibits structural patterns that are qualitatively complex enough to be interesting, but simple enough to be easily quantifiable; though it exhibits structural regularities (syntax), it is free of complications imposed by semantics; it has clear adaptive functions, such as mate attraction and conspecific recognition; it is amenable to comparative analysis in similarly endowed species; and it is amenable to experimental inquiry of the kind not acceptable in human populations (rearing in isolation, etc.). These among other factors make learned birdsong an attractive test case for hypotheses about human language and the cognitive mechanisms that underpin it. Of particular relevance to this thesis is the following basic question: why is birdsong learned?; given the adaptive benefits of the trait, why aren’t these organisms born knowing their songs?5

5Interestingly, these questions are essentially the opposite to those that arise when thinking about the origins of innate linguistic knowledge, but nevertheless lead to concern for the same kind of underlying co-evolutionary dynamics. The evolutionary rationale for linguistic nativism postulates innate knowledge and argues that evolution would build those structures. For birdsong, the absence of innate knowledge is at odds with that same evolutionary reasoning, raising the question: why wouldn’t evolution build innate
This question has inspired a series of models exploring the co-evolution of genetically encoded predispositions for learning, and culturally transmitted song. Early simulations (Lachlan and Slater, 1999) tested an exciting idea known as ‘the cultural trap’ hypothesis. In brief, the hypothesis goes as follows: where coordination on cultural forms is beneficial (as acquiring species-typical song varieties is for birds), gene-culture co-evolution can favour the emergence of neutrality, or disfavour strong cognitive constraints, because, in a population of flexible individuals, a mutant whose bias rules out the acquisition of some cultural variants will face a cost when interacting with individuals whose variants they cannot acquire. The less biased is the rest of the population, the more varied will be their cultural forms, the more this cost will be incurred by the mutant. Though the population as a whole may be less fit (owing to variation in song type among the population), it cannot be invaded by a mutant. Lachlan and Slater’s simulations support the cultural trap hypothesis under a wide range of conditions, but are quite specific to the case of birdsong, consisting in various assumptions about species neighbourhoods, conspecific recognition, and so on.

A more general treatment is given by Lachlan and Feldman (2003), who present a gene-culture co-evolutionary model testing the stability of two genetic variants that bias acquisition among two arbitrarily distinct cultural forms. Their domain of inquiry is still birdsong, though the authors rightly suggest that their findings might also be relevant to the evolution of constraints on language learning. In these models, genetic variant $A$ rules out the acquisition of one cultural variant, while genetic variant $a$ favours neither cultural variant. In other words, evolution can chose between two cognitive variants: one that is entirely neutral, and one that encodes the strongest possible bias. Results are based on analyses of stable points in recursion equations alongside stochastic simulation of the relevant process. Two significant versions of the model are presented.

Model 1 includes an asymmetry between the sexes that is natural in the birdsong case, but would be not applicable to language: males learn songs, females simply recognise songs. In this model stable points exist at fixation frequencies for both genetic types (both cognitive variants can become universal in the population), depending upon parameter values, initial conditions, etc. However, stability analyses show that popula-

---

6I will refer to models of this type, which allow just two cognitive and two cultural variants, as $2 \times 2$ models.
tions in which the neutral allele \((a, \text{encoding no bias})\) is fixed are more stable than are populations who share the biasing allele \((A)\). That is, biased mutants have a harder time invading populations of unbiased learners than would unbiased mutants in biased populations. These conditions favour the evolution of unbiased learners. The authors explain this as a response to the cost incurred by strongly constrained learners who cannot recognise conspecifics attempting to communicate with the ruled-out cultural variant. This line of reasoning has also been cited in support of the argument that UG could not evolve to be maximally constraining (Nowak et al., 2001): I will review the model on which that claim is based in section 2.2.1.

In model 2, the sexual asymmetry is relaxed such that both sexes may acquire and recognise the cultural forms, bringing the assumptions closer into line with the linguistic case. Fitness now depends upon two factors: i) recognising cultural forms (i.e. the capacity to gain fitness benefits by engaging in interactions with other members of the population \(\text{per se}, \text{as in model 1}\), and ii) sharing a cultural form with other members of the population (frequency dependent selection for parity of cultural traits, or, in their terminology, cultural conformity). Evolutionary dynamics depend upon the balance of these two factors: i) favours the generalist, as in model 1, while ii) favours the biasing allele. These models nicely demonstrate how assumptions about fitness can strongly influence our predictions about the evolution of constraints on learning. Model 2, and the assumptions of sex equality and coordination-based fitness, is a considerably better characterisation of the linguistic scenario. On the terms of this model, then, evolution can support nativism: co-evolution can lead to hard innate constraints prohibiting the acquisition of one cultural variant. In this model, frequency dependent selection for cultural coordination among arbitrary forms favours total nativisation of an arbitrary behavioural variant over neutral learning. Of the models reviewed so far, this is by far the most relevant, representing the evolution of substantive prior knowledge in a c-induction problem for a culturally transmitted trait. However, several aspects of this model compromise the application of its conclusions to language.

Sexual asymmetries aside, it is unclear how conspecific recognition influences the models dynamics, and whether this maps straightforwardly on to the communicative functions of language. Furthermore, cultural evolution is instantiated by random post-learning ‘cultural mutations’ between cultural variants: as the authors note, this is a poor approximation for language change. Such errors in the transmission process
represent an approach to learning that is symptomatic of gene-culture co-evolutionary models and, as I will argue shortly, not a good model for human higher cognitive capacities. Most significant, though, is the tight constraint on genetic types: evolution can only choose between total neutrality and total nativisation. In chapter one, I argued that these are just two extremes on a whole continuum of possible biases. While the hard constraint vs. no bias case may be a convenient formalism, it represents an oversimplifying dichotomy no longer seen as an appropriate model of innateness. Even in the case of birdsong, cognitive predispositions are now known to be soft. For example, biases in song learning recover species-typical song structures only over several generations, in the tutees of birds reared in acoustic isolation (Fehér et al., 2009). Isolated birds themselves sing unusual songs. Were these learners subject to a hard acquisitional constraint, that would be an extremely unlikely outcome; were they unbiased learners, species-typical structures would not re-emerge so readily among tutees.

A second hypothesis for the existence of song learning in birds is also worth noting. The Developmental Stress Hypothesis (e.g. Nowicki et al., 1998) explains the learning of song as an honest signal. Song learning in some male songbirds occurs during a developmental period of potentially high stress. A female bird’s fitness may be increased by avoiding mates who experienced high levels of developmental stress. Males can signal a low-stress youth, and therefore gain a mate, by singing songs which could only have been learned in low-stress conditions. The songs of males who experienced high developmental stress will have been learned less faithfully, and therefore stand out to the female as indices of undesirability. Were the song genetically hard-wired rather than learned, no such discrimination could be effected. Ritchie et al. (2008) demonstrate the logical consistency of this argument with a $2 \times 2$ co-evolutionary model similar in spirit to those just discussed. Support for the hypothesis is shown under a number of conditions which lead to stable populations of flexible learners. Ritchie et al.’s model demonstrates an interesting possibility: cognitive plasticity can itself be an adaptation. This hypothesis has also been put forward for language: later in the chapter, I will review a model that supports exactly this conclusion. However, the dynamics of Ritchie et al.’s model, like those of the birdsong models above, have limited implications for the evolution of language-related cognition, for many of the same reasons (e.g. sexual asymmetries, just two cognitive variants). More generally, the assumptions of the Developmental Stress Hypothesis are simply too specific to the case of birdsong.
2.1.3 Boyd and Richerson’s Direct Bias

In section 2.1.1.1, I considered Boyd and Richerson’s co-evolutionary analysis of the emergence of a preference for social learning. Also in *Culture and the Evolutionary Process*, they present a second co-evolutionary model which explores the evolution of a direct bias. Direct bias is Boyd and Richerson’s terminology for aspects of cognition which directly favour individual traits over others. This is exactly the kind of substantive bias generally assumed by nativist philosophies, and linguistic nativism in particular. Several variations of the model are explored, but basic results are obtained in a $3 \times 2$ model\footnote{3 genetic types, 2 cultural types: there exists one biasing allele for each of the two cultural variant, plus the neutral allele.} comparable to those outlined above: learners inherit either an allele that imposes a bias in favour of one or the other cultural variant, or an allele that encodes no bias among alternative traits. In fact, the bias allele does subtly different things in different versions of the model: it can favour one trait over the other directly, or it can bias a learner to acquire whichever trait is favoured by selection, much like guided variation. These variants of the biasing allele might be thought of as specialist and generalist biases respectively. The co-evolutionary analyses ask whether the biasing alleles can become fixed among a population that learns culturally: like the birdsong models considered above, this model directly captures a co-evolutionary process underlying nativisation of a culturally transmitted trait.

Before considering the co-evolutionary analysis, it is worth briefly mentioning the effects of direct bias on a cultural system, since these will be useful aids when interpreting co-evolutionary findings. In short, a direct bias of any magnitude will, over the course of cultural transmission, bring the favoured trait to fixation in the population relatively quickly. Thereafter the alternative trait will never re-emerge: directly biased transmission leads into a cultural sink. I will discuss the consequences of this dynamic in section 2.1.5. It is also worth noting that selection favours the acquisition of particular traits directly, rather than favouring coordination with other learners: this model assumes that trait acquisition is a problem of *n*-induction.

Two significant co-evolutionary findings are relevant. The first concerns evolution of direct bias in a homogeneous environment. In this scenario, Boyd and Richerson show that, if there is any cost to the direct bias (relative to the neutral allele), co-evolution leads to populations of unbiased learners: the biasing allele is removed.
from the population. This happens because: early presence of direct bias in the population causes fixation of the favoured variant on a cultural time-scale, after which there is no cultural variation in the population. The biasing allele is only favoured when it increases the chance that naive learners will acquire the cultural variant favoured by selection. Once this favoured cultural variant dominates, all learners will acquire the favoured variant, and the advantage to the biased allele is gone. If that allele is costly compared to the neutral allele, it will be removed from the population. If it faces no cost, any ratio of biased:unbiased genes in the population is stable. In a homogeneous environment, the biasing genetic variant never gains a long term selective advantage over the neutral alternatives. This finding hangs on two assumptions: 1) biases are costly, and; 2) when learning from cultural models who exhibit the variant favoured by a direct bias, biased and unbiased learners will acquire that variant with the same probability. Assumption 1) is natural and, as has been demonstrated by many gene-culture co-evolutionary models, an important consideration. Assumption 2), I will argue shortly, is inappropriate for language, and does not hold in other models of cultural transmission.

The second co-evolutionary analysis concerns variable environments. In particular, Boyd and Richerson study evolution among two geographically distinct populations in which selection favours opposite cultural variants, and whose members occasionally migrate between the two. Migration between two distinct populations is necessary to maintain cultural variation in this model because of the ‘cultural sink’ effects brought about by direct bias mentioned above. Where the biasing genetic variant is generalist (biases its possessor to acquire whichever variant is favoured by selection in its current environment), co-evolution can lead to its fixation among both populations. Conditions favouring this result are restrictive though. Again, because of the cultural sink effect, migration must be sufficiently common that the cultural variation necessary for the bias to be favoured by selection is maintained. Furthermore, the strength of selection for the biasing allele is proportional to the strength of the bias imposed by that allele. This means that fixation of weaker biases takes longer than fixation of stronger biases. However, we can imagine that in a more realistic scenario, in which biases of graded strength are possible, and perhaps stronger biases evolve from weaker biases, this dynamic may effect a qualitative change in the outcome of co-evolution.

Where the biasing genetic variant is specialist (directly favours a specific cultural
behaviour) rather than generalist, conditions favouring its evolution are even more restrictive. Relative to its ability to invade a population with generalist biases, the neutral allele can invade populations with specialist biases more easily. To be evolvable, the specialist allele must not be costly, and migration rates must be high. On this basis, Boyd and Richerson conclude that the evolution of any specialist direct bias is an unlikely scenario.

2.1.4 Deafness & Sign-languages

I include this short section on co-evolutionary models of sign-language persistence because, although the topic has received comparably little formal analysis, it offers an exemplary, if unusual, test-case for co-evolutionary theory. The case is unusual because deaf individuals obviously possess a hard constraint which effectively rules out the acquisition of spoken languages, in favour of manual modalities. Consider the scenario as a simplified $2 \times 2$ model: two cultural variants exist, being spoken and manual languages, and individuals can inherit either a hearing gene, which for the sake of argument we can class as virtually unbiased between spoken and manual modalities, or a deafness gene, which encodes a strong bias favouring the manual cultural variant. Since most humans acquire and use spoken languages, the newborn deaf individual enters an environment in which her ruled out cultural variant is at fixation, and the deafness gene is very rare. Such an individual would plausibly face a huge fitness disadvantage by deed of being unable to acquire the common cultural variant. The co-evolutionary models reviewed so far would lead us to predict a decline in the frequency of the deafness gene.

Strikingly, the trend is known to be precisely the opposite: the most common genetic cause of deafness, known as the DFNB1 locus, and the prevalence of sign language users and schools, is increasing, fairly rapidly (Gialluisi et al., 2013). General co-evolutionary models have been developed (Feldman and Aoki, 1992; Aoki and Feldman, 1991) that attempt to find conditions which could explain this observation. These models show that the empirically observed trend can be explained if two conditions hold: 1) sign language is transmitted from genetic parent to child, and 2) deaf individuals selectively mate with other deaf individuals. These conditions correspond to vertical cultural transmission and assortative mating (Nance et al., 2000) respectively.

---

There are many: current estimates of more than 120 genes are uncontroversial (Nance, 2003)
Gialluisi et al. (2013) generalise these models to accommodate a distinct class of village sign languages in which both deaf and hearing people can acquire the sign language. Roberts and De Vos (2014) analyse similar models for conditions under which a late-introduced deafness gene could become stable among communities of hearing individuals. Their simulations show that small population sizes and weak biases towards spoken language among the hearing can maintain the gene for several generations. Results from these models are not easily generalisable, owing to topic specific factors such as assortative mating and recessive inheritance. However, the case is a paradigm example that, when cultural transmission is involved, seemingly implausible dynamics can shape cognitive evolution in unpredictable ways: co-evolutionary models explain those dynamics, and show that the details of cultural transmission and bias strength can have radical evolutionary consequences.

2.1.5 General Discussion of Gene-Culture Co-evolutionary Models

This section concludes the review of gene-culture co-evolutionary models. Above all, these models, and the debates that surround them, show that co-evolutionary dynamics can be complex and unpredictable. For various reasons already noted, none of the models reviewed captures the co-evolutionary process that underpins the evolution of innate constraints on language. Here I expand on some of these points, and consider some features of gene-culture co-evolutionary models generally that compromise their application to the question at hand.

2.1.5.1 Technical & Formal Issues

First, I wish to air some known concerns with the general approach to model analysis in this field. These generally relate to practices that reflect issues of mathematical necessity, convenience, or elegance, but might nonetheless be profitably relaxed. One such concern relates to the widespread use of stability analysis on equilibrium models. As noted by Kendal et al. (2009), many models in this field are built on the assumption that, because cultural processes are thought to evolve much more rapidly than do biological systems, the relevant target for natural selection is a cultural system that has...
already converged on its stable state. Often such a state is assumed, and invasion or stability dynamics are studied with respect to genetically homogeneous, culturally stable populations. Analyses of this kind often are unable to capture complex patterns of reciprocal causation: in chapter one, I argued that these are are fundamental to language. On this basis, I will pursue fully dynamic co-evolutionary analyses of genetic and cultural traits in the models I will derive in later chapters. Whatever this approach loses in mathematical elegance, it gains in qualitative insight: early co-evolutionary time-course dynamics will turn out to be of great interest. Another common practice is to focus on the average fitness of a population. As Perreault et al. (2012) suggest, selection generally acts on individuals, not populations, and, as models of social learning have taught us, co-evolutionary systems don’t always respond transparently to fitness measures.

Most significantly, gene-culture co-evolutionary models almost universally\(^{10}\) study populations that contain just two genetic types or cognitive variants. These usually correspond to one unbiasing variant and one variant which imposes the strongest possible bias. Allowing just these extremes rules out an entire continuum of possible biases, as I have argued above. It also obscures any dynamics that might result from the presence of and competition between multiple co-existing biases of different strengths. While it may complicate the mathematics, the analyses I will pursue will allow many graded biological types to simultaneously exist in the population, subject to mutation between types. Evolution shouldn’t be constrained to choose between just two extremes on a continuum of possibilities.

2.1.5.2 An Emphasis on N-Induction

Many models of gene-culture co-evolution assume n-induction rather than c-induction. As I have argued in chapter one, c-induction is fundamental to language (Chater and Christiansen, 2010), and might well induce complex dynamics that affect how inductive biases evolve. Models of the origin of social learning in particular almost universally assume that the target of learning is externally defined. Though there are models that assume c-induction, such as the birdsong models discussed in section 2.1.2, these generally involve topic-specific assumptions. As Chater and Christiansen suggest, c-induction may underpin many human behaviours: basic general models of

\(^{10}\)Though see Perreault et al. (2012) for an exception to the rule.
gene-culture interactions in c-induction problems might well be broadly valuable.

2.1.5.3 Class of Adaptations

Of the gene-culture co-evolutionary models that relate to cognition, by far the most common are those that explore the evolution of social learning biases. As I argued above, these concern cognitive adaptations that don’t encode substantive prior dispositions to acquire particular behaviours. Nativism in general, linguistic nativism in particular, and the associated evolutionary arguments, generally concern substantive prior knowledge, such as preferences for particular linguistic structures (particular constituent ordering patterns, or for hierarchical structures, for example). Gene-culture co-evolutionary models rarely study the evolution of this kind of knowledge: Boyd and Richerson’s direct bias model has received comparatively little attention since its development, for example. Furthermore, many models of gene-culture co-evolution aren’t concerned with cognitive evolution whatsoever: models of lactase persistence or sex ratios are classic examples. Such cases hinge on dynamics that cannot be directly equated with those that underpin the evolution of specialised cognition: cognitive evolution is special, even among other gene-culture co-evolutionary processes, precisely because the capacity being shaped is an information gathering device.\footnote{Ironically, the case of lactase persistence tends to be the go-to example of gene-culture co-evolution in discussions about the origins of language (e.g. Briscoe, 2000; Hurford, 2011; Számadó and Szathmáry, 2012; Bolhuis et al., 2014).}

2.1.5.4 Approaches to Learning

Gene-culture co-evolutionary models generally assume extremely simplified models of learning and cultural transmission. Even the most relevant models, which explore the evolution of substantive cognitive biases for convergence on arbitrary cultural forms (such as the birdsong models of section 2.1.2), tend not to include any substantial model of cognition. It is worth exploring exploring this point in detail, since assumptions about the nature of learning are known to have potentially dramatic consequences for co-evolutionary dynamics (e.g. Mitchener, 2007), and because it relates to the general criticism that models from this field rarely include the fundamentally inferential character of cultural transmission (see e.g. Claidière et al., 2014).

In virtually all gene-culture co-evolutionary models, learners observe traits in full: nothing about the teaching individuals is unobserved. A fundamental property of lan-
guage acquisition, and of higher human cognitive functions more generally, is that the
behaviour a learner observes does not fully specify the trait of interest: there is a dis-
connect between the teacher’s underlying, unobservable mental representation, and
the sparse realisations of that representation observable in behaviour. For example,
the utterances a language learner observes, or the sentences it hears, are noisy, limited
samples from which she must reconstruct an underlying representation. This capacity to
reconstruct a generative model that is fundamentally unobservable in full is central
to nativist reasoning, and is a strong desideratum of any model from which one would
wish to make claims about linguistic nativism. Without this inferential ambiguity, there
exists no real role for the nature of observed information in these models: A learner’s
observations cannot be more or less diagnostic of their generating representation, more
or less variable, greater or fewer, etc\textsuperscript{12}.

Reasoning about the internal state of another individual on the basis of ambiguous
evidence is the essence of human inference about other humans, and an uncontrover-
sial fact of language acquisition. While directly copying traits from others may be an
appropriate model for simpler behaviours (how to hold a table-tennis bat is a classic
example from Boyd and Richerson), or for learning in other species, it is not a desirable
representation of language acquisition, nor, perhaps, of many other human behaviours
of interest (political views is another example from Boyd and Richerson). It is precisely
because the mind is able to arrive at such rich representations from such sparse obser-
vations that nativist philosophies exist.

Boyd and Richerson’s (1985) model of direct bias is a useful example of this in prac-
tice. For instance, imagine a naive learner encounters two cultural parents, both of
whom exhibit behavioural variant 1. In Boyd and Richerson’s model, whether or not
the learner is biased, he will only ever consider the observed variant, which will there-
fore always be induced\textsuperscript{13}. It is central to nativist reasoning about language that, in this
scenario, despite observing behaviourally homogeneous models, there remains some
probability that the unbiased learner also considers variant 2 because, however ho-
mogeneous among models, his observations will always under specify the generating
representation. In these models, inferential ambiguity only occurs when there exists

\textsuperscript{12}Though Perreault et al. (2012) go some way towards this, as do Ritchie et al. (2008).

\textsuperscript{13}Some models allow random post-acquisition mutations in behaviour, such as in the birdsong models
of Lachlan and Feldman (2003), for example. This is a step in the right direction for language, but still
qualitatively oversimplified: errors in language learning are principled, and depend upon the kind and
magnitude of behaviour observed.
between-individual trait variation among the pool of teachers; bias only influences the outcome of trait acquisition when teachers don’t agree. However, in language acquisition, the learner always faces inferential ambiguity, even among consistent models: this is the foundation of the APS.

Casting this hypothetical scenario in linguistic terms is a useful exercise. Let cultural variant 1 correspond to a linear-ordering language, and cultural variant 2 represent a structure-dependent language. The biased learner natively favours the structure-dependent language to some significant extent. Boyd and Richerson’s learner, whose teachers both speak the disfavoured linear-ordering language, will always acquire that variant wholesale despite his affections. It is central to nativist reasoning that in this scenario, because the learner can never be certain about its teachers’ true representation, there is some probability that he will ultimately impose structure-dependency on the language he acquires, whether biased or not. This probability is higher in the presence of bias. This difference is just as important in the opposite cultural scenario, where both teachers speak the structure-dependent language. For the nativist, bias in this scenario allows the learner to arrive at the correct representation where the unbiased learner might not. At least, bias increases the probability that the learner is successful, relative to the chance enjoyed by the unbiased learner. In the limit of this reasoning, behavioural observations are so sparse that the unbiased learner cannot reliably induce the correct behavioural variant. In Boyd and Richerson’s model, populations of unbiased learners can maintain highly skewed trait distributions precisely because such inferential ambiguity never arises. That aspect of Boyd and Richerson’s model accounted for several qualitative co-evolutionary insights: principally, this is why a direct bias is an unlikely outcome of co-evolution. Models of cultural evolution that do include inference over potentially ambiguous data consistently show that some bias is required to maintain cultural universals (Smith, 2004; Griffiths and Kalish, 2007; Kirby et al., 2007). The co-evolutionary consequences of this property may be profound. As such, the insights gained from models of gene-culture co-evolution that do not include this aspect cannot be straightforwardly generalised to behaviours like language.

2.2 Co-evolutionary Models of Language

In this section, I turn to co-evolutionary models designed to explore the evolution of innate linguistic knowledge. The review is in three parts: a series of models by Martin
Nowak and colleagues that tend to concentrate on game-theoretic analyses and evolutionary dynamics (section 2.2.1); a series of models in the Iterated Learning paradigm which focus on the relationship between individual biases and population-level distributions of traits (section 2.2.2); and a series of models by Morten Christiansen, Nick Chater, and colleagues designed to exemplify a logical conflict between cultural and biological evolutionary inheritance streams (section 2.2.3).

2.2.1 The Nowak Series

2.2.1.1 Nowak and Krakauer (1999)

Nowak and Krakauer (1999) present four evolutionary game-theoretic analyses which characterise the evolution of several aspects of language under natural selection for improved communication. The first and second analyses concern the evolution of signal-meaning associations (vocabulary systems). The third and fourth analyses concern the evolution of words from sounds, and of grammatical rules from words respectively. I include these analyses in this review because they represent some of the earliest efforts to characterise the evolutionary dynamics associated with cognitive evolution under natural selection for improved communication. However, they are technically not co-evolutionary models.

For example, the first analysis can be summarised as follows. In this model, though there is natural selection, there is no biological evolution. Agents are all born without any signal-meaning associations, or any bias to acquire any particular set of associations. They acquire their system of meaning-signal associations from their biological parent. Agents whose signal-meaning systems allow them to communicate with a larger proportion of the population leave more offspring, and therefore pass their language on to a larger proportion of the following generation. Selection ensures that over time agents’ associations align, while pressures of communication (encoded in the fitness function) ensure agents align on communicatively useful associations (avoiding synonymy etc.). Given an initial population of learners with randomly sampled associations, natural selection can lead to aligned signal-meaning associations among later populations.

This is a model of cultural evolution enriched with natural selection of cultural variants (fitter individuals do more cultural transmission), but there is no biological evolution: the biological population is uniformly unbiased and unchanging. Alter-
natively, transmission of language form parent to child can be interpreted as genetic transmission of an innately specified vocabulary. In this case, the model includes no learning or cultural transmission, and is therefore better suited to questions concerning the evolution of innate communication systems than to the co-evolution of cognition and language. Either way, these models illuminate the evolutionary dynamics of communication between signal-meaning systems well, but say little about the evolution of substantive prior knowledge for a culturally transmitted behaviour.

2.2.1.2 Nowak et al. (1999)

Nowak et al. (1999) extend the models of Nowak and Krakauer (1999) to include biological evolution among competing learning strategies. Learners in these models inherit a bias which dictates whether they learn language from their biological parent, from randomly chosen members of the previous generation, or from successful (biologically fit) members of the population. Some strategies that combine these dispositions are also considered. These models are similar in spirit to the gene-culture co-evolutionary models of social learning discussed in section 2.1.1: they concern the evolution of constraints on attention to available sources of evidence, rather than substantive prior preferences for certain behaviours over others. Gene-culture co-evolutionary models of deafness and sign-languages (see section 2.1.4) show that complex transmission dynamics can result from such preferences, and can potentially bring about counter-intuitive co-evolutionary outcomes. Nevertheless, the process underlying the evolution of such capacities cannot be assumed identical to that underlying the evolution of substantive innate constraints.

2.2.1.3 Nowak et al. (2001)

Nowak et al. (2001) study replicator dynamics among competing grammars and present two analyses. The first is mechanically comparable to that of Nowak and Krakauer (1999), exploring evolutionary dynamics in a single inheritance stream: a number of grammars can exist in a population, learners acquire a grammar from their biological parent through learning, and leave descendants whose number is proportional to a measure of communicative success. These models build on Nowak and Krakauer (1999) principally by allowing richer assumptions about the process of learning, and highlight several important requirements that must be satisfied by the learning process.
Chapter 2: Co-evolution: A Review of Existing Models

if populations are to converge on communicative systems. Nevertheless, as before, the learning process is homogeneous among the population and does not evolve over time.

The second analysis relaxes this assumption, and illuminates two important aspects of cognitive evolution. First, variation among learners’ periods of learning is allowed: a learner’s innate endowment facilitates his exposure to more or fewer empirical observations, or sentences, from which a grammar will be inferred. Stability analysis show that intermediate learning periods are the evolutionarily stable strategy: if the period is too short, learning is too inaccurate; if the learning period is too long, learners face a cost because reproductive output is compromised. Again, this result resembles findings from the gene-culture co-evolutionary models of social learning discussed in section 2.1.1: it concerns the evolution of constraints on attention to available sources of evidence, not the origins of substantive inductive biases.

The next result is more directly relevant: here the analysis concerns selection on the number of hypotheses entertained by learners during inference, or the magnitude of an innately determined set of allowed grammars. Nowak et al. suggest that "in general, there is selection pressure to reduce [the number of hypotheses entertained]..." (2001, pp.116). The reason for this is intuitively clear: "as [the number of hypotheses entertained] declines, the accuracy of grammar acquisition increases." (Nowak et al., 2001, pp.117). Higher inductive accuracy leads to aligned grammars and improved communication. This is essentially a formal instantiation of Pinker & Bloom’s hypothesis that constraints on variation facilitate coordination and are therefore adaptive. The result is intuitive, and appears to bear out the basic evolutionary rationale for linguistic nativism. However, like many gene-culture co-evolutionary models, this model features a restricted representation of innateness that crucially rules out many important possible forms of cognitive specialisation. I have argued that, to be a fair test of the evolutionary plausibility of linguistic nativism, a co-evolutionary model must make available to evolution a set of cognitive predispositions that can vary arbitrarily, or freely, over possible behaviours. That is, the model should allow any distribution (or as close as possible) over hypotheses, including weak biases, strong biases, and hard constraints: it should allow dispositional innateness.

This flexibility is not possible in Nowak et al.’s model. Here, grammars are constrained to be either: i) ruled out entirely (i.e. subject to a hard constraint that prohibits their acquisition); or ii) present in the learners considerations, but assigned no more or
less prior probability than any of the other grammars considered. If the set of grammars a learner can entertain is taken to be those it is able to represent, then the learners in this model should be considered unbiased, and evolution is acting to reduce the representational capacity of the learners. This interpretation is at odds with the notion that the onset of language is often assumed to have involved an *increase* in representational capacity (e.g. Bolhuis et al., 2014) alongside the evolution of inductive biases that guide learners toward specific grammars (Pinker and Bloom, 1990). If the set of entertained grammars is interpreted as the set of non-zero points in a prior probability measure over possible grammars (i.e. the set of grammars that aren’t ruled out by strong inductive biases), then it already encodes infinitely many hard constraints, and evolution acts only to add more hard constraints, while re-distributing leftover probability mass equally over the non-zero grammars. To draw a formal analogy, the learner’s predispositions are constrained to follow a simple step function over grammars: evolution is able to squash or stretch this function, but nothing else. The qualitative consequence is this: evolution can never mould the learner to favour one grammar over the rest, to any degree\(^{14}\). The desirable model might instead assume a fixed set of hypotheses representable by the learner, and take the evolutionary variable to be an arbitrarily varying measure over this set.

Finally, Nowak et al. (2001) also note that there are circumstances in which selection could push in the opposite direction, to increase the size of the hypothesis space:

"Imagine that someone invents a new advantageous grammatical concept that leads to a modified grammar...that is in...[the larger]...but not in...[the smaller hypothesis space]. In this case, the larger...[hypothesis space]...is favoured." (Nowak et al., 2001, pp. 117)

This logic depends on two conditions: that the new grammar is somehow inherently better than alternatives (functionally superior), and that the new grammar is in the larger but not the smaller hypothesis space. While each of these propositions would require motivation, the logic can be generalised to the observation that: where cultural variation exists within a population, more tightly constrained learners might face a cost by deed of being unable or less likely to acquire some valuable subset of the extant cultural traits. The idea is familiar from the gene-culture co-evolutionary models of

---

\(^{14}\)Except where \(n = 1\), in which case there would be no learning whatsoever.
birdsong discussed in section 2.1.2. To be more convincing, Nowak et al. would have to show how such a scenario could arise given the dynamics of the model. Static analyses of stability and invasion often fail to make demonstrations of this kind, and are in this respect less convincing than dynamic co-evolutionary analyses. Nevertheless, Nowak et al. suggest that the balance of these opposing evolutionary pressures could define a stable search-space size sweet-spot that would not be predictable from either pressure in isolation. This is a valuable observation indeed: the analyses I will present will also depend heavily on a delicate balancing of evolutionary forces that carve out potentially counter-intuitive solutions to adaptive problems\textsuperscript{15}.

2.2.2 The Iterated Learning Series

2.2.2.1 Kirby and Hurford (1997)

Kirby and Hurford (1997) analyse computational simulations which implement a co-evolutionary model designed to explore the interaction between the biological evolution of hard constraints on learning and the cultural evolution of language. In their model learners must learn a language of $n$ features. Each individual inherits a genome that contains $n$ genes: the $i$th gene encodes the learner’s preferences concerning the $i$th feature of the language. Two variants of each linguistic feature exist. Three variants exist for each gene: one encoding a hard constraint (a principle) in favour of each linguistic variant, and one which does not bias acquisition (a parameter). Learners acquire a grammar by learning from data generated by learners from the previous generation: cultural transmission occurs through iterated learning. Learning is represented by the \textit{trigger learning algorithm}. Each datum received from a speaker exemplifies one feature of that speaker’s grammar. If that feature matches the feature currently in the learner’s grammar at this position, nothing happens. If a mismatch occurs, the learner randomly chooses one of the features that is not constrained by a genetic bias and flips that feature to the opposite variant. If the change results in a match, the new grammar is maintained. Otherwise, the old variant is maintained. Fitness is calculated after a specified number of such learning episodes. Two measures contribute to a learner’s fitness score, each determined by a number of post-learning communicative interactions: the proportion of received utterances that match his grammar, and the proportion of

\textsuperscript{15}See Mitchener and Nowak (2003) for an extension to this model that studies further conditions of stability between competing UGs.
utterances produced that matched the interlocutor’s grammar.

In sum, the model implements learning over sparse data, cultural transmission, and frequency-dependent selection for grammatical parity (c-induction). In addition to these elements the model includes functional differences between linguistic variants: some linguistic variants are assumed functionally superior to others, and therefore afford the learner an amplified fitness advantage. In the paper, this functionality is related to *parsability*: for some features of the language, linguistic variant 1 is assumed easier to parse than is linguistic variant 2. Increased parsability makes variant 1 easier to acquire during learning and to understand during communication. Kirby and Hurford are chiefly concerned with the emergence of innate constraints that fix *functionally superior* principles among the population.

Two significant analyses are reported. First discussed are co-evolutionary simulations in which functional asymmetries affect communication but not learning; a learner’s fitness is influenced by the parsability of sentences generated by his grammar (and the grammars of individuals he communicates with), but that parsability measure does not influence the uptake of variants during language acquisition. This ensures that the only way functionally superior variants can become fixed in the population is through natural selection of genomes which bias acquisition in the relevant direction. Interestingly, Kirby and Hurford find that despite this selective pressure, evolution doesn’t lead to the fixation of constraints that enforce acquisition of the functionally superior variants. Biasing genetic types become fixed for some but not all features of the language, though where nativisation is observed, this doesn’t seem to reflect the functional pressures. This raises two questions: 1) why aren’t *all* features of the language nativised?; 2) why don’t common genetic constraints favour the functionally superior variants? A candidate response to 1) is that unbiased learning is fairly accurate in this model, and becomes more accurate as more features are nativised. Learners each engage in hundreds of learning trials. Since all learning episodes involve trial and error experimentation with the unbiased features of the language, where these are fewer in number (because some genetic biases have become fixed) learning will be more accurate. One might imagine a threshold, determined by the learning algorithm, the size of the language, and the number of learning trials, after which genetic biasing ensures coordination on enough features that the available learning resources are sufficient to
ensure coordination on the rest\textsuperscript{16}. Question 2) might be resolved by observing that functional concerns make up only a small part of the fitness function: in the simulation reported, the major part of an individual’s fitness score is determined by the requirement for arbitrary parity (c-induction). If by chance the functionally inferior variants become genetically fixed under this pressure, mutants biased to acquire the more functional but less common variant will not gain an advantage.

The second analysis allows functional asymmetries to influence acquisition as well as use. In this case biases favouring the functionally superior variants become fixed in the population. Cultural transmission ensures the more functional variants quickly take over, after which genetic biases favouring their acquisition are fixed by natural selection. This dynamic is comparable to that observed in Boyd and Richerson’s (1985) models, in which cultural processes rapidly stabilise beneficial variants. In their models, no subsequent nativisation was observed because, when learning from a homogeneous cultural pool, unbiased learners uniformly acquire the common trait. In section 2.1 I argued that this property of the model compromised the application of its conclusions to language. Kirby and Hurford’s simulations support this observation. Their model is mechanically very similar to Boyd and Richerson’s: the major difference is that, in Kirby and Hurford’s model, the unbiased learner is not guaranteed to acquire the language of its teachers. Genetic biases therefore remain valuable after cultural convergence, and nativisation of the functional variants ensues. Kirby and Hurford conclude that evolution does support linguistic nativism, but only if cultural transmission can stabilise a subset of linguistic variants for biological evolution to track.

Several aspects of the model qualify the generality of this conclusion. Again, the constraint that evolution can only choose between hard constraints or total neutrality is chief among these: dispositional innateness is off the table. The model also represents quite specific assumptions about the spatial arrangement of populations. More generally, the model includes a fairly specific model of language acquisition that makes a range of potentially confounding assumptions: for example, randomly re-sampling a linguistic principle in a randomly chosen part of the grammar when the learner encounters data that are inconsistent with her hypothesis. Where gene-culture co-evolutionary models included too few assumptions about cognition, these simulations include too many: general conclusions require a principled but minimal model of learning.

\textsuperscript{16}Also see Harvey (1993) and Wiles et al. (2001) for general discussions of why genetic algorithms often stop short of total nativisation.
Notwithstanding these features, Kirby and Hurford’s simulations demonstrate a valuable insight of general significance that will also feature heavily in the models I derive: cultural processes can distort the fitness landscape explored by biological evolution. In other words, cultural evolution can make adaptive cognitive variants that might otherwise not be favoured by selection. This hints at a broad possibility, which I will argue for strongly in later chapters, that rather than prohibit adaptation (e.g. Christiansen and Chater, 2008), culture might in fact licence, and even encourage, certain classes of cognitive adaptation.

2.2.2.2 Briscoe (2000)

Briscoe (2000) presents a series of simulations which explore how properties of a language acquisition device influence the cultural evolution of language. Among a host of pioneering analyses, Briscoe includes two co-evolutionary models designed to directly explore issues of reciprocal interaction in the evolution of language related cognition. Briscoe’s model is particularly relevant because it focuses on the evolution of inductive biases: it includes an explicit distinction between hard, inviolable constraints and soft, defeasible biases. The analysis is extremely rich, consisting in extensive treatments of language acquisition and use. This depth uncovers a number of interesting questions, but also means the results are difficult to interpret closely, and limits the generality of the model’s dynamics. For this reason, and for reasons of space, here I’ll simply summarise the basics of the co-evolutionary model, and discuss the key results. Full details of all analyses can be found in Briscoe (2000).

In this model, agents each possess a UG that specifies principles and parameters for a generalised categorial grammar: principles represent inviolable hard constraints; parameters represent soft constraints that can be initially unset (unbiasing) or set to some default value (biasing). A default inheritance network specifies a partial ordering of principles and parameters (and associated atomic categories, etc.) in the grammar. Language acquisition involves parsing strings received from other learners, and updating parameters on the basis of whatever derivations the parser returns for those strings. Each learner is equipped with a parser that uses a deterministic, bounded-context shift-reduce algorithm. The parser is also enriched with an algorithm to compute working memory load. Parameter updating is based on the trigger learning algorithm, and enriched with various extras such that it suits the default inheritance network, accounts...
for limited memory, etc.

Each agent has an age which determines a critical period for language acquisition, a reproductive window, and ultimately the agent’s death. Populations go through interaction cycles of dynamic length (number of interactions), in which younger agents learn from other randomly selected agents. Population turnover is gradual: older members die and younger members reproduce at the end of interaction cycles. Reproduction occurs via sexual recombination though single-point crossover and point mutation (which transforms individual principles into unset or default parameters, and vice versa). Each agent’s fitness is based on the success of its interactions, and is based on several costs and benefits that result from factors such as parsing ease, memory load, and the number of updated parameters, etc. Crucially, fitness is partly determined by successful communication with other agents: the model implements frequency-dependent selection for linguistic parity. Computational simulations of the process assume finite populations and a specified number of interaction cycles.

Two analyses of the model are reported. The evolutionary variables of most relevance here are the relative proportions of principles, default-valued parameters, and unset parameters. In the first analysis, Briscoe (2000) shows that over time, the population frequency of default-valued parameters increases at the expense of unset parameters, showing that "[the language acquisition device]...is evolving an accurate language-specific inductive bias..." (Briscoe, 2000, pp. 284). The population frequency of principles remains roughly constant, begging the question: why are default-valued parameters (soft inductive biases) favoured over principles (hard inviolable constraints)? The second analysis introduces migration: adult agents are sporadically inserted into the population, leading to greater linguistic variation. This scenario also leads to the evolution of innate constraints: the population frequency of unset parameters decreases, being replaced by principles and default parameters in roughly equal number. This represents "...genetic assimilation of both inductive biases (defaults) and hard constraints (principles)..." (Briscoe, 2000, pp. 286). Crucially: "The replacement of unset parameters by principles is an example of the type of genetic assimilation Pinker and Bloom (1990) envisage..." (Briscoe, 2000, pp. 286). However, an analysis of populations that are initially all subject to hard constraints shows that those constraints are partially eroded over time. This result parallels the findings of Nowak et al. (2001): evolution seems to

---

17This analysis in particular is a direct response to Deacon’s (1997) argument - which prefigure the moving target argument - that rapid language change will prohibit cognitive adaptation.
favour some nativisation, but not total nativisation.

Precisely why hard constraints proliferate (in addition to inductive biases) in the second analysis but not the first is unclear. Briscoe offers a novel explanation for the evolution of innate constraints in the face of cultural evolution: in a finite population, even in the context of rapid language change, only a finite subset of the (potentially vast) space of possible languages will ever be explored; a learner who is constrained to favour that subset will always gain an advantage. Cultural evolution can be rapid, but it can never explore the entire space of possible languages. Briscoe’s analyses demonstrate that co-evolutionary dynamics can lead to innate constraints, but that it is far from clear what kind of innate constraints will be favoured where several are possible. He argues that soft constraints can play the same role as hard constraints in practice, but concludes that "...further experiments are needed to explore the degree of genetic assimilation of principles as opposed to default parameters..." (Briscoe, 2000, pp. 288).

The model presents a comprehensive demonstration that cognitive specialisation is possible but complicated in the context of culture: it includes reciprocal interactions, c-induction, and several kinds of innateness. However, the myriad of specific assumptions obscures the core elements of the co-evolutionary dynamic: to understand the basic relationship between culture, cognition, and evolution, a more general analysis that isolates the fundamentals of the process is required.

2.2.2.3 Smith (2004)

Smith (2004) explores the co-evolution of shared vocabulary systems and cognitive biases that guide their acquisition. In particular Smith is concerned with the evolution of biases for one-to-one mappings between meanings and signal. Such a bias is known to guide child language acquisition, and may be responsible for much of the regularity observed in lexical systems. I situate Smith’s models in the Iterated Learning series since they focus strongly on the dynamics of iterated learning, but they might be equally well understood as extensions to the Nowak series just reviewed in section 2.2.1. Many similarities link the two. Most importantly, though, both concern the acquisition and evolution of meaning-signal association matrices. Smith’s analyses reveal a number of important insights, both substantial and methodological. In particular, they focus on first understanding the dynamics of cultural transmission given the language model assumed, only then asking how these dynamics might complicate biological evolution,
Chapter 2: Co-evolution: A Review of Existing Models

and relating these issues to the relative merits of static and dynamic co-evolutionary analyses.

Three types of learner are studied: those biased for \((h_+)\) and against \((h_-)\) acquiring vocabulary systems that include homonyms, and those who are unbiased with respect to homonymy \((h?)\). Fitness in these models reflects communicative accuracy, and depends upon both parity and functional asymmetries among languages (vocabulary systems that do not include homonyms allow more successful communication and higher fitness payoffs when shared among learners). Gains are accrued for resolving references in the utterances of other learners (understanding), and for producing utterances from which other learners can infer the correct reference (being understood). Initial analyses show that long-term cultural transmission among cognitively homogeneous populations results in shared vocabulary systems that are: communicatively optimal among \(h_-\) learners; characteristically homonymous, and therefore communicatively sub-optimal, among \(h_+\) learners, and; partially homonymous among \(h?\) learners.

Next, the pairwise evolvability of these learner types are examined: in a population that is genetically homogeneous in one type, how likely is a mutant of another type to reproduce and spread? Both static and dynamic analyses are presented, and the differences between these are crucial. Static analyses illuminate the mutant’s fate if she is born into an otherwise genetically homogeneous population whose culture has already converged to a stable state. As discussed in section 2.1.5.1, this approach is common among gene-culture co-evolutionary analyses, and is often justified by the suggestion that the speed of cultural change dramatically exceeds that of biological change. Smith shows that both \(h_-\) and \(h?\) populations are evolutionarily stable: they resist invasion by mutants with different biases. It is clear why \(h_-\) populations resist invasion: learners biased to acquire communicatively optimal vocabularies will gain the maximum fitness advantage, while mutant \(h?\) or \(h_+\) learners will, through bias or through noise, tend to impose some degree of homonymy on the systems they acquire, and therefore face fitness disadvantages and eventually die out. However, the particularly relevant question is: why can’t \(h_-\) learners invade populations of \(h?\) learners? Of the scenarios considered, this situation arguably enjoys greatest ecological validity: it is conservative to assume in these models that at some point a bias of interest arose among otherwise unbiased learners, and must have been able to spread in these conditions.

The answer lies in the fact that cultural transmission has already converged to sys-
tems that reflect the $h_?$ variant. Some parts of the shared vocabulary include homonyms: $h_-$ learners will be unable or less likely to learn these parts of the system, and therefore acquire some meaning-signal associations that are not shared with other learners. Relative to the $h_?$ learners, the $h_-$ mutant will have lower communicative success and therefore eventually die out. This appears to seriously complicate the evolutionary account of the origin of the mutual-exclusivity bias. If it cannot invade a population of unbiased learners, how could it arise? Resolving this question requires a dynamic analysis.

Smith presents two dynamic analyses in which the initial populations are made up by 90% of one type of learner, and 10% of the mutant type. The first analysis assumes that a shared vocabulary system has already emerged at the beginning of these simulations. The second analysis assumes random initial vocabularies. Crucially: both allow the cultural system at subsequent generations to be influenced by the genetic biases present in the population. Rather than assume a converged stable cultural system that reflects only the biases of the indigenous learners, the dynamic analysis allows that mutants shape the vocabulary system too. This modification has several substantial consequences, but the most relevant for my purposes concerns the maintenance of $h_-$ mutants among $h_?$ populations. In this case, small sub-populations of $h_-$ learners can persist through genetic drift and are able to converge on shared vocabularies that are more functional (less homonymous) than those used by $h_?$ learners. This offsets the fitness costs they incur through being unable to resolve ambiguous utterances sometimes spoken by $h_?$ learners, and allows $h_-$ learners to gradually take over the population.

This is a direct consequence of reciprocal interactions between genes and culture. The signature of this process is captured by the dynamic analysis but not by the static analysis, and has lead to opposing conclusions between techniques. Loosely, the dynamic analysis shows that evolution of a domain-specific anti-homonymy bias is possible, while the static analysis rules out this possibility.

Like Kirby and Hurford’s model, this model revolves around deep functional asymmetries between languages, and allows only three cognitive variants. The representation of biological evolution does not allow mutation between those types or any costs to specialisation, and the analysis only considers pairwise comparisons among cognitive types. For these reasons, among others, it leaves many questions about the evolution of inductive biases unanswered. However, it provides a quintessential demonstration
of the following principle: cultural transmission under the influence of evolving biases leads to a reciprocal interaction which may not be captured by static analyses of evolutionary stability. Cultural evolution is cumulative, and its trajectory is sensitive to the introduction of mutant learners. It is potentially not safe to draw conclusions about the evolution of language-related cognition on the basis of pre-hoc assumptions about the relative speeds and independence of cultural and biological inheritance streams. Smith’s analyses show that relaxing such assumptions can turn their logic upside down: if biological change is slower than cultural change, genetic drift or weak selection may facilitate periods of biological stability sufficient in duration to allow biological variants not currently favoured by selection to impose significant effects on cultural systems. Such changes in turn might modify future selection differentials. Dynamics of this kind are potentially rich and potentially counter-intuitive, and are best studied through fully dynamic co-evolutionary analyses.

2.2.2.4 Smith and Kirby (2008)

Smith and Kirby (2008) derive a co-evolutionary model designed to explore the co-evolution of languages and the prior pre-dispositions of Bayesian language learners. Of all the models I will review in this chapter, Smith and Kirby’s is by far the most relevant: the models I will derive should be seen as extensions in this stream of research. Following Kirby et al. (2007), Smith and Kirby assume the task facing learners consists in inducing a system of associations that relate each of $n$ meanings with one of $m$ signal classes. Learners must induce such a system on the basis of observing sets of meaning-signal pairings generated by previous learners. The Bayesian inferential model allows precise specification of learners’ prior expectations. In this model, those expectations describe a preference for regularity in the association of meanings and signals: learners are biased in favour of languages that employ fewer signal classes across possible meanings. The strength of this bias is controlled by a parameter $\alpha$ whose range is effectively unbounded, allowing arbitrarily weak ($\alpha \to +\infty$) or strong ($\alpha \to 1$) biases towards regular languages. During inference, Bayesian learners combine a likelihood distribution implied by the observed data with a prior distribution which reflects innate pre-dispositions, in order to obtain a posterior distribution which reflects the learner’s relative degree of confidence that her teacher was speaking each of the conceivable languages. Having computed the posterior distribution, the learner must use this in-
formation to induce a single language. Following Griffiths and Kalish (2007), Smith and Kirby allow two types of learner that correspond to differing strategies for this task: sample learners, who use posterior probabilities to randomly sample a language, and; MAP learners, who induce the language with highest posterior probability.

Cultural transmission among homogeneous populations of learners is known to follow different trajectories between these two types of learner (Griffiths and Kalish, 2007). In populations of sample learners, long term cultural transmission converges to a stationary distribution over languages that is isomorphic to the distribution imposed by the learners’ prior preferences: weak biases lead to variable cultural systems, while only hard constraints can lead to a cultural universal. In populations of MAP learners, learners’ prior preferences are amplified over the course of long-term transmission: biases of any strength can ensure favoured cultural variants eventually dominate. The degree to which learners’ biases are amplified depends upon aspects of the transmission process: for instance, where transmission occurs through observations that are more sparse, greater amplification results. Smith and Kirby assume that selection favours arbitrary coordination among learners, as measured by an individual’s ability to communicate with the her sub-population (whose members have converged on a single grammar). Through stability analyses, two questions are studied: 1) which biases are evolutionarily stable? and; 2) which hypothesis selection strategy is evolutionarily stable? Several findings emerge.

Methodologically, Smith and Kirby assume a genetically homogeneous population of learners whose culture has converged to a stable state, and compute the fate of a mutant learner born into this population. A finite discretised range for $\alpha$ is assumed, and pairwise invasion dynamics are explored between different values of $\alpha$. For populations of sample learners, selection tends to favour lower values of $\alpha$, or stronger biases, as predicted by the adaptationist rationale for linguistic nativism. However, the strength of this dynamic is dependent upon the nature of the incumbent biases:

The degree to which stronger priors are favoured in sampling populations is somewhat sensitive to the strength of the population prior, however. For example, populations with a very weak prior bias in favour of regularity ($\alpha = 40$) resist invasion by mutants with much stronger prior preferences for regularity ($\alpha = 1$)...this tendency for weaker majority priors to reduce the extent to which strengthened priors can invade also pertains for populations where the majority have a flat, unbiased prior: in such populations every
language is equally probable, and any bias to acquire a particular language is penalized due to the consequent decreased ability to acquire the other languages. Consequently, there are two evolutionarily stable strategies in sampling populations: the strongest possible prior ($\alpha = 1$) or a completely unbiased prior. (Smith and Kirby, 2008, pp. 3599)

The logic is familiar from models discussed in sections 2.1.2, 2.2.1, and 2.2.2.3: innate constraints can be disfavoured (in conditions of cultural variation) because they rule out communication with speakers of the disfavoured language. The analysis also implies a principal of general significance that I will discuss at length in later sections: though a given configuration of a co-evolutionary system might be stable against invasion, or maximise fitness, it may not be reachable given the dynamics of the system. Stability does not imply inevitability.

Smith & Kirby’s analysis suggests that a population of neutral learners could never be invaded by biased mutants, and therefore could never evolve biases of any kind despite their advantages. However, a subsequent analysis relaxes the assumption that communicative accuracy is judged only within a sub-population, and instead assumes a population-wide measure of communicative accuracy. Under this regime, the unbiased prior is not evolutionarily stable: evolution in populations of sample learners does favour nativisation. These are exactly the kinds of complication that were shown by Smith (2004) to be impenetrable under static stability analyses, and which may evade informal reasoning. Smith and Kirby’s analysis, which is limited to pairwise comparisons among culturally stable homogeneous populations, is suggestive but inconclusive. Only a dynamic analysis, which allowed arbitrarily heterogeneous populations, reciprocal interactions, and mutation between biological types could resolve questions about the evolution of innate constraints in initially unbiased populations of sample learners.

Evolutionary dynamics in populations of MAP learners differ markedly. Smith and Kirby show wide ranging selective neutrality over bias strength: since weak biases are amplified, their effects are equivalent to stronger biases, and they consequently afford learners equivalent fitness benefits. As predicted informally by Kirby et al. (2007), bias strength is shielded from selection. On this basis Smith and Kirby conclude that in MAP learner populations evolution favours the biased learner, but does not favour the construction of stronger biases. This dynamic will also feature heavily in the models I derive. Though its consequences for nativist theories of language may be profound, I
defer their discussion until later chapters.

Two additional results are reported. First, Smith and Kirby assess the evolutionary stability of the two strategies (MAP and sample learners) when competing. They find that MAP learning is favoured, since the MAP learner will tend to acquire a favoured language with greater probability than would the equivalently biased sample learner. Second, Smith and Kirby demonstrate that if learners face a cost proportional to the strength of their bias, evolution will favour the weaker biases in MAP populations. Taken together these results suggest a complex evolutionary dynamic: evolution favours a learning device whose particulars become largely invisible to selection, and which may consequently be vulnerable to degradational pressures that chip away at its information content. Smith and Kirby conclude that this evolutionary dynamic compromises cognitive adaptation to language:

We would not be surprised if species-specific specializations for the acquisition of linguistic structure turn out to be rare or even non-existent – this is what the evolutionary argument...suggests. (Smith and Kirby, 2008, pp.3601)

The models I derive will support a similar evolutionary pattern: bias amplification, selective neutrality, and degradation. However, the analyses I will report ultimately lead to a different conclusion: the subtle interplay between biology and culture can provide a rapid and robust route to cognitive specialisation. Smith and Kirby’s model is, of all those reviewed in this chapter, the best suited to test the evolutionary rationale for nativism: it identifies rich and surprising dynamics that suggest the evolution of language specific cognition might not be so straightforward.

2.2.3 The C&C Series

2.2.3.1 Chater et al. (2009)

Chater et al. (2009) present an evolutionary analysis widely thought to demonstrate that in principle, the evolution of domain-specific biases for arbitrary properties of language is impossible. Results formalise the moving target hypothesis:

"...as the processes of language change are much more rapid than processes of genetic change, language constitutes a "moving target" both over time and across different human populations, and, hence, cannot provide a stable
environment to which language genes could have adapted...a biologically determined UG is not evolutionarily viable. " (Christiansen and Chater, 2008, pp.489)

"...the linguistic environment over which selectional pressures operate presents a "moving target" for natural selection. If linguistic conventions change more rapidly than genes change via natural selection, then genes that encode biases for particular conventions will be eliminated - because, as the language changes, the biases will be incorrect, and, hence, decrease fitness."(Christiansen and Chater, 2008, pp.489)

Three versions of the model are tested, each of which I will consider in detail individually, but the basic structure of the model is as follows. There exists in the population a language which possesses \( n \) independent principles. Each principle takes one of two arbitrarily distinct variants which, for consistency with previous models, I will refer to as variants 1 and 2 of the \( i \)th principal. Each learner inherits a genome consisting of \( n \) genes. The \( i \)th gene determines the learner’s predispositions concerning the \( i \)th feature of the language. For each gene there are three possible alleles: one neutral, one which biases the learner towards variant 1, and one which biases the learner towards variant 2. For each of the \( n \) principles, the learner’s task is to induce whichever variant the extant language specifies. Learning consists in trial and error sampling of variants. For each principle, the learner must repeatedly sample one of the two variants until the correct variant is chosen. Sampling probabilities for the \( i \)th principle are determined by the \( i \)th gene. The neutral allele imposes a uniform distribution over variants, so each is equally likely to be sampled. The biasing alleles weight sampling in favour of the corresponding variant, such that the favoured variant is sampled with probability \( p \), and the disfavoured variant with probability \( 1 - p \), where \( 0.8 \leq p \leq 1 \) is a parameter of the simulation. A learner’s fitness is inversely proportional to the number of trials required to acquire the correct variant for each principle. During reproduction the newborn individual inherits half of its genes from one parent and half from the other. Each gene is subject to mutation at some small probability, in which case that gene would be randomly re-assigned one of the three relevant alleles.

The first analysis presented is a proof of principle demonstration that the model can bring about the biological evolution of innate predispositions. Principles of the language are all fixed at variant 1, and under various parameter settings and starting conditions, Chater et al. demonstrate through simulation in finite populations that over
time, the population’s genes come to reflect the language: all learners eventually possess the relevant biasing allele at all genes. Note that this is not a co-evolutionary model. In fact, the model is best seen as multiple parallel but independent instantiations of a simple optimisation algorithm: all principles of the language are independent of each other and of the learners, are arbitrarily distinct, remain fixed at the same value throughout, and are the same for all $m$ members of the population; all genes are independent of each other, arbitrarily distinct, and mutate among alleles with symmetric probabilities; and all members of the population are selected according to the same criteria. In this sense, the simulation is a little like evolving a single gene that can match variant 1, 2, or neither, to match variant 1 $m \times n$ times in parallel. Nevertheless, the basic evolutionary point is demonstrated convincingly: evolution leads to nativisation in a stable environment.

The second analysis presented introduces language change, and instantiates the moving target idea. Language change is implemented by modifying a randomly chosen principle of the language at each generation, with some probability $l$ that is a parameter of the simulation. The main finding here is that, as $l$ increases, the long term frequency of the biasing alleles decreases: language change prevents the biological evolution of substantive biases; selection favours generalist, or unbiased learners.

Before discussing the third analysis, it will be helpful to review some key aspects of the model. First, it is worth pointing out that this result is, notwithstanding some technical differences, a replication of Boyd and Richerson’s (1985) finding that variable environments prevent the evolution of direct bias. In fact the models are very similar. If we overlook the fact that Chater et al. evolve $n$ genes/principles in parallel, and concentrate on the mechanics of evolution and learning, several parallels are clear: both models allow two (opposing) biasing alleles and one neutral; both assume an externally defined target for learning (n-induction); both assume the learning target fluctuates randomly; and both select learners according to their probability of acquiring the correct variant. The major difference between the two is that Boyd and Richerson’s model implements cultural transmission, while Chater et al.’s does not. Boyd and Richerson’s learners use information gleaned from others around them to acquire a trait, or information directly observed from the environment in similar models (e.g. guided variation), while Chater et al.’s learners simply sample traits according to their innate predispositions.

The third analysis reported in this paper concerns a modification to the manner in
which the principles of the extant language are determined. As noted above, preceeding versions of this model assumed n-induction: the language to be learned is entirely divorced from learners’ dispositions. Here that premise is relaxed, and a parameter is added to the model which allows learners’ biases to influence linguistic principles:

...for the ith principle at...[generation]...t + 1, there is a probability, g, that this principle is determined by the genetic allele that is most prevalent at the ith location at [generation] t. Otherwise, with probability 1 - g, the principle is not influenced by the genes, but is subject to linguistic change with probability l, as before. (Chater et al., 2009, pp. 1017)

The parameter $g$ effectively controls the degree of reciprocal causation. Where $g = 1$, the genetic make-up of the population fully determines which variants are favoured by selection, implementing a kind of frequency dependent selection for genetic parity: the evolutionary ‘goal’ is that genes coordinate. Where $g = 0$ this is exactly the same simulation as in the previous analysis, and the evolutionary ‘goal’ is that genes adapt to an externally defined language. The main result is that, roughly, where $g \geq 0.5$, selection favours biasing alleles and the neutral variants are wiped out: evolution leads to hard-wired biases for linguistic principles.

To interpret these results, it is useful to acknowledge again that, while the language and the genomes contain $n$ principles and genes, these are independent and structurally identical, and as such are subject to identical evolutionary dynamics, licensing us to imagine $n = 1$ without qualitative loss. In that case, this model is very similar to Lachlan and Feldman’s (2003) birdsong model 2 review in section 2.1.2: both assume frequency dependent selection among genes that can encode either neutral or strongly biased substantive prior knowledge. Again, the main difference between these models is that the birdsong case includes (rudimentary) cultural transmission, where the language model does not.

Overall, Chater et al.’s results suggest the following: 1) evolutionary dynamics in n-induction problems with stable selective regularities lead to strongly constrained learners; 2) evolutionary dynamics in n-induction problems with variable selective pressures lead to unconstrained learners; and 3) evolutionary dynamics that include reciprocal causation lead to strongly constrained learners.

Point 1) is at odds with Boyd and Richerson’s finding that evolution of direct bias in stable environments leads to generalist learners. This is because Boyd and Richer-
son’s model included cultural transmission and learning, and these forces were able to ensure learners acquired the correct variant without the need for genetic biasing. The comparison highlights an important point: the addition of cultural transmission in otherwise similar models can change the outcome of co-evolution radically. Point 2) reflects the main claim of the paper, and relies on the logic that in variable environments it doesn’t pay to be genetically committed to some cultural variant that is likely to disappear; variable environments favour generalist learners. The finding is familiar from gene-culture co-evolutionary models, and bears out the moving target hypothesis as applied to language. Point 3) is perhaps the most relevant. Chater et al. in fact dismiss this result on various grounds, which are worth briefly unpicking. First, of the fully-genetically-determined case they say:

If $g$ is 1, then the language is "reinvented" afresh at each generation to fit the genetic biases in the population. In [this] case...the linguistic properties are already determined by preexisting genes. (Chater et al., 2009, pp.1017)

And for intermediate values of $g$:

[This scenario] seems implausible for most, if not all, arbitrary linguistic principles; if selection pressure on the relevant locus is influenced mainly by the genes, then this indicates that the principle has a strong preexisting genetic basis. (Chater et al., 2009, pp.1017)

As Chater et al. say, these scenarios seem implausible: language is not re-invented afresh each generation, nor does selection act directly over the genetic underpinnings of language-specific cognition. The implausibility of these conditions results from the nature of the model, which doesn’t include cultural learning or any distinction between genotype and phenotype. The plausible equivalent scenario, in which selection favours coordination among the languages ultimately induced by the learners, but those languages are learned from other individuals under the influence of genetic biases, remains to be tested.

In summary, Chater et al.’s main claim, stated technically, is that: in rapidly varying problems of n-induction, selection favours neutral organisms over organisms constrained to acquire just one possible trait. This is a valuable observation of general significance, and is in accord with similar models from related fields. However, various features of the model mean that it does not capture the co-evolutionary process I
have argued underlies the evolutionary rationale for nativism. Like many other models, innateness is restricted to neutrality or strong constraints, and does not feature costs to specialisation. More fundamentally, it does not capture the process of cultural transmission or allow c-induction, which are crucial features of the evolutionary dynamics I aim to examine.

2.2.3.2 Christiansen et al. (2011)

The models under consideration by Chater et al. (2009) explored adaptation for functionally arbitrary features of language. In this paper, Christiansen et al. build on those earlier models to test evolutionary dynamics among what they term functional aspects of language. Their conclusion based on this model is that evolution of strong innate constraints is possible in this scenario: evolution supports nativism about functional aspects of language. At this point it is worth clearing up some terminology, since mine is at odds with Christiansen et al.’s. In my usage of the terminology, to describe some linguistic variant as functionally arbitrary is to say that it is not inherently better or worse than alternative variants; the function it serves can be satisfied equally well by any variant. In the evolutionary terms relevant to these models of language evolution, being functionally arbitrary means: no variant inherently accords the learner a larger fitness advantage than does any other variant, all else being equal. Conversely, to have an inherent functional advantage means that, all else being equal, some variants are favoured by selection.

For example, vocabulary is, for the most part, functionally arbitrary: whether a linguistic community describes a four-wheeled motor vehicle as a car or a voiture makes no difference, so long as the community agrees on a form. This reflects the need for arbitrary parity that Pinker and Bloom (1990) claimed is central to the evolution of strong innate constraints. Alternatively, whether that community agrees to use their vocal apparatus or their feet to express signals would make a difference: making the latter choice would plausibly engender a fitness disadvantage, since being able to walk and talk at the same time is useful; the vocal apparatus is functionally superior to feet. Crucially, the functionality of a variant is distinct from how widely it is shared in the population. To measure the fitness value of a variant with respect to how widely it is shared in the population is to assume frequency-dependent selection that results from the need for coordination. Function and frequency are two related but separate concerns that
arise when thinking about the fitness benefits that result from communication. In my terminology, then, Christiansen et al.’s model does not include functional asymmetry. Rather, fitness depends upon arbitrary coordination with other learners. Language acquisition is treated as a problem of c-induction, and as such the model implements frequency-dependent selection for linguistic parity, bringing it closer in line with the co-evolutionary process I wish to explore.

The basic structure of the model is identical to that of Chater et al. (2009): \( n \) linguistic principles must be learned, each of which can take one of two possible forms; learners possess \( n \) genes, one for each principle, each of which biases the learner strongly toward one or the other variant, or is neutral. Agents in these simulations learn their language by communicating. At each generation, each agent interacts with other randomly selected members of the population. Interactions are deemed successful if, among all the principles in both agents’ languages, most share the same variant. If the shared majority variant is more common in agent A’s language than in agent B’s, agent B resamples a variant for one of the principles about which the two disagreed. Several details of this model of transmission and learning qualify the generality of its dynamics. With respect to transmission, it assures agents only learn from conspecifics whose language is similar to their own. With respect to learning, agents still re-sample variants on the basis of their genetic dispositions alone, rather than perform inference over data generated by the other agent. The fitness function rewards parity, but also includes asymmetries in the benefits gained through production and comprehension\(^{18}\).

In sum, the analysis shows that, in a c-induction problem with no empirical learning but parity-conditional interaction, evolution can fix strong biases among the population. It builds on earlier models in this series by allowing interaction and coordination-based fitness, and shows that these assumptions can overcome the moving target problem in this framework. However, it also shares limitations with those models (e.g. no dispositional innateness, rudimentary learning model), and represents the dynamics of a more specific, rather than a more general, set of assumptions.

2.2.3.3 Baronchelli et al. (2012)

Baronchelli et al. (2012) re-analyse Chater et al.’s (2009) model under new population

\(^{18}\)Note that this model is also very similar to that of Kirby and Hurford (1997), but does not include the same kind of functional asymmetries between languages.
dynamics. The paper, titled *The Biological Origin of Linguistic Diversity*, aims to find evolutionary scenarios that are consistent with two observations: A) languages vary enormously, yet B) any child is able to learn any language (equally easily). Four scenarios are considered. Each involves population splitting: roughly half way through the simulations, the population of agents is split into two geographically distinct sub-populations.

The first analysis replicates Chater et al. (2009)’s main set of simulations (analysis 2, n-induction with language change) in these conditions. Chater et al. (2009)’s results are, naturally, recovered: when language change is low, evolution equips both populations with strong genetically-determined biases that match the local language, which differs between populations. This scenario is consistent with observation A), but not with B), since any given learner is hard-wired to learn only his local language. Where language change is high, neutral genes predominate in both populations, and language varies between the populations, consistent with both A) and B).

The second analysis allows probabilistic migration between the sub-populations, such that an agent’s fitness is determined by its ability to learn both languages. Even under low rates of languages change, Baronchelli et al. show that any significant probability of migrating, and therefore of having to learn the other language, puts the brakes on cognitive specialisation: neutral genes predominate in both populations. Again, this is consistent with both A) and B). Note that this analysis is an even closer replication of Boyd and Richerson’s (1985) model of the evolution of direct bias in a varying environment, which also allowed population splitting and cross-migration. The third analysis is analogous to Chater et al.’s final simulations which allowed the target of learning to be influenced by the genetic makeup of the learners. Where genes influence language strongly, language change is overruled and strong biases are fixed among both populations. This scenario is deemed implausible because either: 1) populations diverge genetically and linguistically, contradicting observation B); or 2) populations do not diverge genetically or linguistically, contradicting observation A).

The final analysis allows variation in the rate of language change. In particular, Baronchelli et al. test the scenario in which language initially changes slowly but later begins to change rapidly. Genetic biases predominate early in the simulations, but are then replaced by neutral alternatives as language change accelerates. Again, this parallels Boyd and Richerson’s model of direct bias evolution: biases can emerge and then
later be lost or degraded as their utility erodes. Baronchelli et al. interpret plasticity as an adaptive outcome: where before neutrality was explained as a consequence of restrictions on the rate of genetic evolution (e.g. Chater et al., 2009), now it is discussed in adaptationist terms:

"Only biological adaptations for flexible learning combined with cultural evolution can explain how each child has the potential to learn any human language" (Baronchelli et al., 2012, pp.1)

"...our results point to an evolved genetic predisposition to accommodate the continual cultural evolution of language". (Baronchelli et al., 2012, pp. 5)

Of course the two kinds of explanation for plasticity aren’t mutually exclusive. But it is worth noting that these models don’t allow an alternative factor that might also favour the generalist more parsimoniously: cost to specialisation. Restrictions of the model paint Baronchelli et al. into a corner. Genes can encode only a hard constraint or a neutral learner: thus, plasticity is only possible through neutrality. If biased learners can never acquire disfavoured language types efficiently, they can never be favoured in variable environments. If the only alternative is total neutrality, then the authors are forced into the conclusion that evolution cannot bias language acquisition whatsoever. Many of my arguments in later sections of this thesis will depend upon the idea that, in cultural systems of learning, it is possible to be substantively biased in such a way that does not prohibit plasticity yet does guarantee the acquisition of locally common variants. Dynamics of this kind, which I will suggest are precisely what makes gene-culture co-evolution so rich, are off the table in this model, leading to an explanatory trap.

In sum, based on evolutionary considerations, Baronchelli et al. claim that the language learner’s ability to acquire any language, or individual plasticity, implies of the supporting cognitive apparatus three properties: 1) it is universal and uniform among the species; 2) it does not feature language-specific biases, and; 3) it is an adaptation for the ability to learn many languages.

2.2.3.4 Baronchelli et al. (2013)

Baronchelli et al. (2013) present a co-evolutionary model designed to explore adaptation in environments that vary. Qualitatively, no significant new findings are presented:
results demonstrate again that, where the target of adaptation is externally defined and varies independently of the organisms adapting to it, selection favours generalists. The value of this model, as Baronchelli et al. note, is in the generality and tractability of the formal framework. Theirs is a relatively lightweight mathematical framework based on a stochastic interacting particle model from statistical physics. The model is analytically tractable at the mean field level: this permits broader coverage of parameter space and a fuller understanding of the effects of individual parameters.

Since these results replicate those discussed earlier, and again the model concerns n-induction rather than c-induction, I will not review its details in depth. However, there are at least two related features of the model that represent interesting additions to the series, and are therefore worth noting for inclusion in the models I will derive: namely, studying classes of genetic types rather than individual types, and allowing asymmetric mutation between these classes as a way to impose principled costs to specialisation. This is a small but significant addition to the representation of innateness, which has been a limiting factor in many models reviewed in this chapter.

Three classes of learner\(^{19}\) are considered: specialists, generalists, and maladapted learners. We have learned from gene-culture co-evolutionary models that cost can play an important role in cognitive evolution. Studying classes of learners is a useful way to account for the cost of specialisation in a principled way through mutation dynamics among classes. For example, Baronchelli et al. make the assumption that mutation among generalists is class-neutral and tends to result in a generalist mutant, while mutation among specialists is generally deleterious and therefore results in a mutant that belongs to the less fit generalist class. Degradational mutation is empirically motivated (Sanjuán et al., 2004; Eyre-Walker and Keightley, 2007), and represents a principled way of imposing an evolutionary cost to cognitive specialisation. This and other aspects of the model are formally admirable, but the analysis reveals little that has not already been discussed in this review, and is subject to roughly the same limitations as were the models of Chater et al. (2009).

This concludes the review of co-evolutionary models designed to explore the evolution of innate linguistic knowledge. Many of these models reveal crucial insights into

---

\(^{19}\) Though terminologically I discuss this model as relating to cognitive evolution among biased learners, it is presented in the paper primarily as a domain neutral model which applies most obviously to problems like species adaptation to rapid climate change.
co-evolutionary dynamics that underpin the evolution of language-specific constraints on cognition. For instance, models from the Nowak series (section 2.2.1) provide a robust formalisation of Pinker and Bloom’s basic argument: pressures of communication favour the evolution of constraints on cognition. Models from the Iterated Learning series (section 2.2.2) suggest that culture can distort the fitness landscape implied by those pressures, leading to counter-intuitive cognitive outcomes. Models from the C&C series (section 2.2.3) question the capacity of these forces to mitigate against rapid cultural evolution. Together they present a powerful case that the evolutionary rationale for linguistic nativism should be reconsidered in the light of culture. However, individually, none comprehensively describes the process under study in this thesis. Despite an expansive literature of related models, the basic dynamics that characterise the evolution of inductive bias in the context of cultural transmission remain obscure: there exists no general, minimal model of the dynamic process underpinning the evolutionary rationale for linguistic nativism. In the models I will derive, I aim to combine the merits of the different existing approaches to the topic: the formal precision of models in the Nowak series; the centre stage given to cognition and transmission by models in the Iterated Learning series; and the simplicity and abstractness of models in the C&C series.

2.3 Summary

In summary, the models of gene-culture co-evolution reviewed in section 2.1 generally benefit from a formal precision, abstractness, and simplicity that affords clear interpretation of their dynamics. However, they tend not to include much cognition, and are often analysed with techniques not ideally suited to the reciprocal interactions so central to my concerns: they are designed to answer subtly different questions. Conversely, the language evolution models reviewed in section 2.2 generally include richer treatments of cognition and cultural transmission, but are often complex, intractable models in which it is difficult to isolate the basic principles that govern the interactions between cognition, culture, and evolution. Both literatures would benefit from a general model of the co-evolutionary process that underpins the evolution of inductive biases for behaviours that are, like language, culturally transmitted through expression, inference, and induction. The reviews in this and the previous chapter have revealed a number of desiderata that, in order to be a comprehensive analysis of the evolutionary
rationale for linguistic nativism, a model of this process must meet. I bring this chapter to a close by explicitly summarising these requirements below.

2.3.1 Model Desiderata

- **D1: Many graded biological types.** -
  Since Locke, our conception of innateness has admitted a continuum of innate knowledge ranging from presence at birth (hard constraint), through dispositional innateness (strong & weak biases), to pure empiricism (neutral learners). A fair model of the evolution of innate linguistic knowledge should not rule out swathes of this continuum.

- **D2: C-Induction.** -
  Though functional differences and learner-external pressures are important considerations, these must be distinguished from pressures that result from the requirement for arbitrary parity. The desirable model should allow learners to learn from data generated by other learners; it should assume the goal of learning is to arrive at coordinated representations, and allow the frequency-dependent selection that results.

- **D3: Evolve substantive prior knowledge.** -
  The model must allow the evolution of preferences that favour or disfavour specific representations directly; the principal evolutionary variable must be substantive inductive bias.

- **D4: Statistical inference over sparse data.** -
  The model must include a principled model of learning: it should include a distinction between mental representations and public expression of data portraying that representation; it must allow the nativist premise that those public expressions (linguistic data) never fully specify the underlying representation, leading to inferential ambiguity (the poverty of the stimulus); it must reflect the contemporary view that language acquisition involves induction through principled statistical inference that can act over structured representations (and be guided by inductive bias).

- **D5: Fully dynamic co-evolutionary analysis.** -
  The desirable co-evolutionary analysis will be fully dynamic: it will allow many
biological types to exist in a population *simultaneously*, and allow principled mutation among those types; it will allow the behaviour of the model to determine the constitution of the population over which selection acts; it will make no *a priori* assumptions about the relative speeds of the two inheritance streams, and allow each to influence the other dynamically in real time.

- **D6: Costs to specialisation.** -

  The model must account for proportional costs to specialisation. The most convincing demonstration that co-evolutionary dynamics favour the evolution of strong inductive biases would show that the process includes but overcomes costs to such specialisation.

- **D7: Minimal & General language model.** -

  The model should not include dynamics that are overly specific to any one grammar formalism or model of language; it should be as simple as is possible while still maintaining the qualitative properties of interest; it should be directly comparable to gene-culture co-evolutionary models of other behaviours.

In the next chapter, I take these desiderata as a starting point, and derive and analyse a mathematical model that captures the co-evolutionary dynamics underpinning the evolution of innate inductive biases for culturally transmitted behaviours.
Chapter 3

The Evolution of Inductive Biases for Culture: A Formal Analysis

This chapter presents a co-evolutionary model that describes the evolution of domain-specific inductive bias in the context of culture. In section 3.1 I derive a mathematical model of gene-culture co-evolution which captures each of the desiderata identified in the previous chapter. Section 3.2 presents the results of several analyses of the model, under a range of assumptions. Finally, I discuss the model’s dynamics, and the implications they entail for nativist theories, in section 3.3.

3.1 A Co-evolutionary Model

This section presents details of a mathematical model of gene-culture co-evolution. The model is built from three components: models of language learning (section 3.1.1), biological evolution (section 3.1.2), and cultural transmission (section 3.1.3). Once these are defined, a formal framework to integrate the individual components and analyse the resulting co-evolutionary system’s dynamics is presented in section 3.1.4.

3.1.1 A Model of Language Learning

3.1.1.1 Preliminaries

The model of language learning and production should satisfy the following requirements. First, it should be as simple as possible whilst still exhibiting the qualitative
properties of interest. It will distinguish between just two possible representational states, each of which is associated with a typical pattern of behaviour. In linguistically familiar terms, representational states are grammars and behavioural patterns are utterances. The model will allow two possible grammars, each of which is associated with its own diagnostic utterance, but which can also occasionally generate the utterance ordinarily associated with the competing language. Second, the model of learning should capture the contemporary view that language acquisition is grounded in principled statistical inference over structured representations that can be guided by inductive bias. The model should be able to represent arbitrarily weak or strong inductive biases, and the strength of those inductive biases should be clearly distinguishable. It should operate at the computational level of analysis, such that it is not tied to any specific representation of the computational procedures it performs. Third, the cognitive model should be sufficiently general and flexible that: 1) it can be used in analyses of experimental data in later chapters, and; 2) it can be directly compared to models of cognition employed in the wider relevant literature, such that it is clear how the predictions of the evolutionary model relate to existing claims about innateness. Finally, the model should lead to cultural evolutionary dynamics that are well understood. In particular, it will be extremely useful to understand how inductive bias in the model relates to the long-term population-level distribution of behaviours that results from cultural transmission among learners of this kind. Each of these desiderata can be met well by a Bayesian probabilistic model of statistical inference acting over a hypothesis space of two simple probabilistic context-free grammars (PCFGs). Before presenting the formal details of this model (section 3.1.1.3), I’ll ground the learning problem in an illustrative linguistic example.

3.1.1.2 A Grounded Example: Word-Order Constraints

Though the mechanics of the language model are general, it is useful to ground it in a concrete linguistic example. Nativist reasoning about language is arguably strongest in the domain of syntax:

"From the beginning of generative grammar, syntax was seen as the distinctive characteristic of language - the component that provides language with its creative properties, and the one in which resides the lion’s share of complexity and abstractness. It has been in terms of syntax that the cen-
tral problems of acquisition and innateness have been framed and argued” (Culicover and Jackendoff, 2005, pp.530)

In particular, decades of theoretical, experimental, and typological research in linguistics has been dedicated to illuminating the origins of regularities in the typological distribution of constituent ordering patterns (Greenberg, 1963). The word-order universals, which describe typologically common constraints on the relations between elements of a clause, have been the subject of intense inquiry, including extensive typological surveying (e.g. Dryer, 1992), behavioural experiments (e.g. Cook, 1988), formal theory (e.g. Chomsky, 1965), and, more recently, phylogenetic analysis (e.g. Dunn et al., 2011; Maurits and Griffiths, 2014); they have been central to classic linguistic nativism (see e.g. Chomsky and Lasnik, 1993). Word-order universals are a particularly appealing example in which to ground the model because, in its contemporary form, the debate surrounding their origins is an example par excellence of present-day approaches to universals. On the one hand, word-order universals may be brought about by cultural or historical factors independent of the cognitive particulars of language learners (see Dunn et al., 2011, for example). On the other, those same universals may be brought about by, and therefore reflect, biases in the cognitive capacities of language learners:

"A small number of the logically possible word order configurations account for a large proportion of actual human languages. To explain this distribution, typologists often invoke principles of human cognition which might make certain orders easier or harder to learn or use." (Tily et al., 2011b, pp. 1364)

Those biases may include language-specific substantive innate constraints (see e.g. Culbertson et al., 2012), or be grounded in domain-general cognitive capacities (e.g. Christiansen, 2000). The nativist case for the centrality of substantive language-specific biases in the origin of word-order universals has been given an explicit evolutionary rationale (Pinker and Bloom, 1990). Most importantly, though, there exist Bayesian cognitive models designed to capture the acquisition of grammars that constrain the use of competing word-orders under the influence of innate biases. For instance, Culbertson et al. (2012) and Culbertson and Smolensky (2012) derive a Bayesian model of grammar induction that acts over probabilistic context-free grammars (hereafter PCFGs) which impose probabilistic constraints on the use of harmonic vs. disharmonic word-order
patterns. Crucially, that model allows arbitrary prior beliefs about the possible ordering patterns. Those beliefs can be understood to reflect shared innate domain-specific biases that impose constraints on what will be acquired during learning, and are directly comparable to inductive biases in models of other linguistic phenomena (see e.g. Clark and Lappin, 2010).

In sum, word-order patterns are a classic and contemporary subject of nativist reasoning about language, have been linked to an adaptive rationale, and have been modelled with simple, well-understood grammar formalisms¹ and inferential models that meet all the requirements I have outlined above. I’ll derive the model with notation that reflects this example as a way to ground these issues in a familiar linguistic domain, but note that the fundamentals of the model are general: they could apply equally well to any linguistic feature, or wider learned behaviour, that is acquired in the same way. Neither the particular linguistic domain, nor the specific grammar formalism I will adopt, are crucial to the model: for example, nothing substantial rests on the fact the grammars are context-free. The crucial aspects of the model are: it allows a set of mental representations; each representation is probabilistically associated with particular expressions of behaviour; these representations can be reverse engineered through statistical inference; and inference can be guided by inductive bias that directly favours particular representations. Though I’ll dress-up this foundation as PCFG induction for a constituent-ordering problem, it assumes nothing more than statistical inference over simple statistical grammars.

3.1.1.3 A Bayesian Model of PCFG Induction

While it is mechanically similar to, for instance, Culbertson and Smolensky’s (2012) model, the model abstracts over particular instances of constituent ordering patterns, and instead assumes just two anonymous grammatical categories, X and Y. The grammars that learners must infer dictate constraints on the ordering of these elements. Let $L = (N, T, S, R, \theta)$ represent a generic PCFG framework in Chomsky normal form: $N = \{S, X, Y\}$ is the set of non-terminal symbols; $T$ the set of terminal symbols; $S \in N$ the Start symbol; $R$ a set of rules defining possible expansions of LHS non-terminals of the form $A \rightarrow B_1B_2 \ldots B_k$ where $A \in N$, $B_i \in N \cup T$, $i = 1, \ldots, k$ and $k \geq 0$; and

¹See Clark and Lappin (2010), in particular chapter 10, for a formal discussion of the use of PCFGs and statistical inference models in nativist theorising about language.
\( \bar{\theta} \) a vector of probabilities for rules in \( R \) such that \( \theta_{A \rightarrow B} \in \bar{\theta} \) gives the probability of expansion \( A \rightarrow B, |\bar{\theta}| = |R| \), and for all non-terminals \( A \in N \), and corresponding rules \( A \rightarrow B \in R \), relevant rule probabilities sum to unity so that \( \sum_{A \rightarrow B} \theta_{A \rightarrow B} = 1 \). Within this framework, the model allows just two possible grammars: \( L_0 = (N,T,S,R,\bar{\theta}_0) \) and \( L_1 = (N,T,S,R,\bar{\theta}_1) \). In both of these grammars, the expansions are as follows:\(^2\):

\[
S \rightarrow XY \\
S \rightarrow YX \\
X \rightarrow ... \\
Y \rightarrow ...
\]

Both grammars allow expansion of \( S \) to both possible orderings of anonymous grammatical categories \( X \) and \( Y \). The (terminal) right hand sides of \( X \) and \( Y \) can be ignored:\(^3\):

I assume components of utterances to be labelled for grammatical category. That is, the learner is able to recognise grammatical categories \( X \) and \( Y \) in the terminal strings produced by these PCFGs, and is reasoning exclusively about the underlying grammatical constraints on relative orderings of \( X \) and \( Y \). Each grammar is associated with a diagnostic but not deterministic utterance: a dominant ordering pattern for \( X \) and \( Y \). Therefore, the grammars contain fixed mirror image rule probabilities, so that:

\[
\bar{\theta}_0 = (\theta_{S \rightarrow XY} = p, \theta_{S \rightarrow YX} = 1 - p) \\
\bar{\theta}_1 = (\theta_{S \rightarrow XY} = 1 - p, \theta_{S \rightarrow YX} = p)
\]

Rule probabilities are fixed in this way so that, where \( p \neq 0.5 \), each grammar favours a distinct ordering pattern. If \( p = 1 \), these would be mutually exclusive grammars with zero overlap in productions, each exclusively employing a single ordering pattern, thus essentially reducing to CFGs, each with a single distinct rule for expanding \( S \). Where

\(^2\)I have chosen to leave \( X \) and \( Y \) as anonymous grammatical categories with anonymous expansions for two main reasons: to avoid the reading that I am making claims about specific linguistic phenomena, and to highlight the generality of the model.

\(^3\)The RHSs of \( X \) and \( Y \) include no non-terminals, and are distinct (non-overlapping). An equivalent alternative approach would be to treat \( X \) and \( Y \) as terminals. I have chosen to treat them as non-terminals to be intuitively in line with the notion that they can be understood as grammatical categories, rather than specific strings.
0.5 < p < 1, each grammar favours a distinct ordering pattern yet permits expansion of S to both possible orderings of X and Y. This is the situation I instantiate here: utterances should be indicative of a particular grammar but not deterministically diagnostic, such that the learner will face inferential ambiguity during language acquisition. I assume the learner knows p: this corresponds to the assumption that the learner entertains two possible grammars that represent simple complementary alternatives, each favouring a different ordering pattern. The inferential task facing the language learner is therefore one of PCFG induction: given an observed set of utterances, the learner must reason about which of these two grammars generated those utterances and on the basis of these calculations induce one or the other PCFG. I assume simple statistical inference as follows.

Assume the learner has observed N productions from one of these grammars, and has identified among those utterances y applications of the rule S → X Y, with y ∈ 0, 1, ..., N. Since there are only two re-write rules for S, it follows that the count of applications of the rule S → Y X is given by N − y, so we can focus exclusively on the former count by keeping track of y7. In the rational framework, the learner is assumed to compute posterior probabilities P(Ll|y, N) and P(Ll|y, N) for each possible grammar. To do this the learner must combine the likelihood of those particular observations if the hypothesised grammar were correct, P(y|Ll), l = 0, 1, and the prior probability of that grammar P(Ll). Learners combine these sources of information in accordance with the principles of Bayesian inference so that:

$$P(L_l|y) = \frac{P(y|L_l)P(L_l)}{P(y)},$$

(3.4)

where P(y) is a normalising factor giving the probability of observations y averaged over all hypotheses, and can be ignored for now. Assuming each of the N utterances are statistically independent and identically distributed, the likelihood of observing exactly y applications of the rule S → X Y, given a total of N productions and known

Footnotes:

4 Note that this is also true for the equivalent scenario in which 0 < p < 0.5.

5 The alternative assumption that the learner infers p would specify a continuous hypothesis space of grammars. I will analyse a model with exactly this assumption in the next chapter.

6 I have re-assigned upper-case N to refer to the number of observed utterances rather than the set of non-terminals in a given PCFG. This is to bring the following notation into line with common practice - I will not need to refer to the set of non-terminals again.

7 These counts are sufficient statistics for the model of inference. Working with these counts as summaries of the data simplifies calculations later.
rule probabilities $\bar{\theta}_0$ and $\bar{\theta}_1$, is determined by the Binomial sampling scheme, so that, for the respective grammars $L_0$ and $L_1$:

$$
P(y|L_0) = p^y (1 - p)^{N-y} \quad (3.5)$$

$$
P(y|L_1) = (1 - p)^y p^{N-y} \quad (3.6)$$

The prior probability of a given grammar, $P(L_l)$, represents the innate inductive bias of the learner: by favouring one or the other grammar, it imposes substantive probabilistic constraints on the ordering pattern the learner is disposed to acquire. I assume the simplest possible prior distribution over grammars, so that $P(L_1) = \alpha$ and $P(L_0) = 1 - \alpha$, where $\alpha \in [0, 1]$. Where $\alpha = 0.5$, both grammars are equally probable and the learner is unbiased. Extreme values $\alpha \approx 1$ and $\alpha \approx 0$ impose hard constraints on what the learner can acquire, ruling out the acquisition of grammars $L_0$ and $L_1$ respectively. Intermediate values of $\alpha$ impose weaker biases in favour of one or the other grammar.

Having computed the posterior distribution over grammars, the learners must then induce a single grammar. Here I consider two approaches to this task, or hypothesis selection strategies. In this Bayesian framework, hypothesis selection strategies can have dramatic impacts on the long term cultural dynamics induced by the model of learning (Griffiths and Kalish, 2007). Firstly, a learner might draw a random sample from the posterior distribution, so that $P_{\text{sample}}(L_l|y) = P(L_l|y)$. I will refer to this type of learner as a sample learner. This strategy does not give rise to the cultural amplification effect discussed in the previous chapters. Secondly, a learner might simply induce the maximum a posteriori (MAP) grammar, which is the grammar that is accorded the highest posterior probability, so that:

$$
P_{\text{MAP}}(L_l) = \begin{cases} 
1, & \text{if } P(L_l|y) > P(L_{l'}|y) \\
0.5, & \text{if } P(L_l|y) = P(L_{l'}|y) \\
0, & \text{otherwise} 
\end{cases} \quad (3.7)$$

I will refer to this type of learner as a MAP learner. This strategy does lead to the amplification effect often associated with cultural transmission. Throughout this thesis I will, following others (e.g. Griffiths and Kalish, 2007; Kirby et al., 2007; Smith and Kirby, 2008), contrast these two types of learner as proxies for the two different kinds
of cultural dynamic they give rise to: those that do and do not lead to amplification.

It will in later sections become convenient to work with posterior odds rather than posterior probabilities. Since there are only two competing hypothesis, and for each the denominator in (3.4) is identical, we can simply combine prior odds and likelihood ratios to gain posterior odds:

\[
O_{L_0} = \frac{P(L_0|y)}{P(L_1|y)} = \frac{1 - \alpha}{\alpha} \times \frac{P(y|L_0)}{P(y|L_1)} = \frac{1 - \alpha}{\alpha} \frac{C(N, y) p^y (1 - p)^{N-y}}{C(N, y) p^{y-N} (1 - p)^{N-y}} = \frac{1 - \alpha}{\alpha} \left( \frac{p}{1 - p} \right)^{2y-N},
\]

and likewise:

\[
O_{L_1} = \frac{P(L_1|y)}{P(L_0|y)} = \frac{\alpha}{1 - \alpha} \left( \frac{1 - p}{p} \right)^{2y-N},
\]

where \(C(N, y) = \binom{N}{y}\) represents the binomial coefficient. This is just another approach to assessing posterior probabilities for competing grammars, and will be useful mathematically, but does not change the model. Finally, production of utterances follows the Binomial sampling scheme already outlined: once a learner has induced a grammar and is called upon to generate productions, those utterances follow the Binomial likelihoods outlined in equations (3.5) and (3.6). Thus, production probabilities and the likelihood calculations a learner has access to during inference are identical: when reverse engineering the behaviour of others, the learner consults her own linguistic model\(^8\).

This completes the model of language learning. There are three free parameters relevant to the following analysis: \(N\), the size of the dataset a learner observes; \(p\), which dictates how strongly grammars are associated with a particular ordering pattern, and; whether learners are MAP learners or sample learners. With this model, I have arguably assumed a great deal of the learner: that she can recognise the relevant grammatical

---

\(^8\)Previous models of Bayesian iterated learning include a noise term in production to ensure ambiguity in the data. Here \(p\) performs that role (where \(p \neq 0, 1\)).
categories in strings, and that she knows the rule probabilities of the grammars, for example. This is a deliberate strategy to concentrate the evolutionary analysis on one central variable: substantive inductive bias. Nevertheless, to be clear about the central assumptions: rationally combining prior knowledge and evidence from observations corresponds to a domain general capacity for statistical inference; that the learner can represent the PCFGs corresponds to the capacity to represent structured knowledge in this domain; that \( p \) is fixed rather than learned represents a constraint on the space of hypotheses the learner is able to represent; the prior distribution, given by \( \alpha \), that can favour one PCFG over the other corresponds to the domain-specific innate inductive bias: the evolutionary rationale for nativism predicts that evolution will lead to a strongly skewed prior distribution (\( \alpha \approx 0, \alpha \approx 1 \)). The prior distribution will be the principal subject of the following analyses. To explore the evolution of learners’ innate biases, \( \alpha \), requires a model of biological inheritance through which learners come to be biased.

### 3.1.2 A Model of Biological Inheritance

#### 3.1.2.1 Preliminaries

The model of biological inheritance should satisfy the following initial requirements: i) the space of possible biological types must be finite, since the evolutionary analysis I will present in section 3.1.4 requires this; ii) mutation between types must degrade the information content of the prior, such that it imposes costs proportional to specialisation. Mutation should also specify a computable distribution over types that can act as a baseline which would result from neutral drift, and; iii) there must be sufficient biological types to encode a wide range of possible prior distributions. Put simply, I require a smooth and structured space of biological types that maps nicely onto the parameter range of the prior and penalises strongly biasing genomes. Each of these requirements is met well by the following bit-string genome representation. Bit-string formalisms are a common way to represent innate information in evolutionary analyses (see e.g. Langton, 1995), and have been widely used in language evolution models (e.g. Kirby and Hurford, 1997; Chater et al., 2009).
3.1.2.2 Genome Model

Assume each learner inherits from its biological parent a genome which determines its prior, or its innate bias. A genome \( q = (q_1, q_2, \ldots, q_n) \) is simply a collection of \( n \) loci, each of which is occupied by a gene from the set \( G = \{0, 1\} \). The set of all possible genomes, \( G^n = \{(q_1, q_2, \ldots, q_n) : q_i \in G \forall i = 1, \ldots, n\} \), is therefore finite and contains \( 2^n \) variants. Recall that the parameter range for priors in this model is \( \alpha \in [0, 1] \). Each genome \( q \in G^n \) maps nicely into this space on the basis of the proportion of its genes taking the value 1, as laid out in (3.10):

\[
\alpha_q = \frac{1}{n} \sum_{i=1}^{n} q_i .
\]  

(3.10)

Though the principal purpose of this genome model is to meet the requirements outlined above, it also has a natural theoretical interpretation. Mapping from genes to priors in this way can be likened to the notion of a polygenic trait that results from the additive effects of many genes. Many human traits are known to be polygenic, and thus exhibit gradation: height, skin colour, hair colour, and eye colour are classic examples of morphological polygenic traits, and general intelligence is the polygenic psychological trait par excellence (Davies et al., 2011). Recent research suggests that traits of this kind are extremely common: large numbers of small-effect loci may be the primary engine of phenotypic evolution (Rockman, 2012). If \( \alpha \) is a psychological trait, each gene in the genome contributes a small effect in favour of one language or the other. Ordering among genes is unimportant for \( \alpha \). For instance, consider the genomes \( p = (0, 0, 0, 1, 1, 1) \) and \( q = (1, 1, 1, 0, 0, 0) \). Though at every locus these genomes differ, both specify the neutral prior \( \alpha_p = \alpha_q = 0.5 \) by deed of sharing an identical distribution of 0s and 1s. This highlights a crucial observation: the mapping from genomes into priors is potentially many-to-one in nature; for a given value of \( \alpha \), there may exist a group of associated genomes. In turn this raises a second crucial factor: the size of these groups varies; some priors may be associated with more genomes than are other priors. It is this property of the model that makes strongly biasing genomes rare. If we were to lay out all possible genomes in \( G^n \), it would be trivial to count how many

\footnote{Note that, despite the formal resemblance, my usage of the many-genes genome representation differs from common practice in language evolution models. For example, in the models of Chater et al. (2009), each individual gene provides information about a separate feature of the language. Here, all genes contribute information to one feature.}
contain \( n - 1 \) 1s, or \( n - 2 \) 1s, for instance. The same count could be performed for all \( n + 1 \) possible groups (genomes can contain 0, 1, 2, \ldots \) 1s), and the relative sizes of these groups would be clear. Fortunately, the Binomial theorem specifies the size of these groups analytically. With some additional formalisation, it is possible to characterise the genome space in a way that allows us to concentrate on meaningful genetic variation. First, let \((G^n, \rho)\) be a metric space, with

\[
\rho(p, q) = \sum_{i=1}^{n} |p_i - q_i|, \quad \text{for any } p, q \in G^n. \tag{3.11}
\]

Since we are only interested in what prior a genome encodes, we only need to know the distribution of genes in that genome. This quantity can be characterised, for any genome \( q \in G^n \), arbitrarily as the number of genes taking the value 1, denoted by \(|q|\), such that \(|q| = \rho(q, 0)\), where 0 is understood as the element \((0, \ldots, 0) \in G^n\). With this we can specify an equivalence relation \( \sim \) on \( G^n \), defined by the rule \( p \sim q \iff |p| = |q| \), which collects all the genomes that specify a given prior into an equivalence class, \( i \in G^n / \sim \), where \( i = \{ q \in G^n : |q| = i \} \), and \( G^n / \sim \) denotes the quotient set of \( G^n \) by \( \sim \). Then, using the binomial theorem, we can be explicit about the size of \( i \), or the number of genomes that specify a given prior, by stating that:

\[
\#i = \binom{n}{i} = \frac{n!}{i!(n-i)!}, \tag{3.12}
\]

With this the space of genomes is partitioned into a manageable \( n + 1 \) groups, each of which specifies a single prior and has a cardinality that can be easily computed. Crucially, the combinatorics of the genome space dictate that groups whose genomes have higher entropy, or less information content, have larger size. That is, there are many more possible genomes that contain \( n/2 \) 1s than there are genomes that contain \( n \) 1s or 0s. Thinking in terms of how genome groups map into priors, this represents exactly the assumption I have appealed to above: the neutral prior, \( \alpha = 0.5 \), is a priori much more probable than biased priors, which become more improbable with distance from neutral. That is, if a learner were assigned a genome from this space at random, it would be much more likely to end up with a genome that encodes a neutral prior than with a genome that encodes a hard constraint. In the full evolutionary model, learners do not inherit genomes at random, of course. There inheritance is determined by selection, reproduction, and mutation. It is the mutation regime that instantiates
the required asymmetries between genome groups, ensuring that stronger priors are costly: they are less likely to emerge through mutation, and more difficult to maintain against mutation pressure.

### 3.1.2.3 Mutation

During inheritance, individual genomes are subject to point mutation at rate $\mu$. Where a mutation does occur, the gene at that locus flips to the opposite gene ($0 \rightarrow 1 \& 1 \rightarrow 0$), so that for any parent $p$ and child $q$ in $G^n$, $P(q_i = p_i) = 1 - \mu$, and $P(q_i \neq p_i) = \mu$. Accordingly, drawing on the metric defined in (3.11), the probability that individual genome $p$ will during reproduction mutate into genome $q$ is given by (3.13):

$$P(\text{child} = q | \text{parent} = p) = \mu^{\rho(p,q)}(1 - \mu)^{n - \rho(p,q)} . \quad (3.13)$$

Whilst this specifies the probability of mutation from one genome into another, it does not give us the probability that, during inheritance, a genome that belongs to class $i$ will mutate into a genome that belongs to class $j$. This question is central to my analysis because it will determine the dynamics of movement around the space of priors through mutation; it determines the probability that the offspring of a learner with bias $\alpha$ will be born with bias $\alpha'$. Thus, an $n + 1 \times n + 1$ matrix $M$ is required, whose components $m_{ij}$ give the probability that a genome belonging to $i$ will mutate into a genome belonging to $j$. Formally, entries in the matrix are:

$$m_{ij} = P(q \in j | p \in i) = P(|q| = j ||p| = i) . \quad (3.14)$$

This quantity is in theory computable by simply counting through all genomes in classes $j \& i$, calculating the pairwise mutation probability for each and weighting the sum by the inverse of the size of the parent class $i$, as in the first line of (3.15). However, where $n$ is larger, the space of possible genomes becomes so large that counting through these sums becomes computationally intractable. Using the binomial theorem, it can be shown\(^\text{10}\) that the computational cost of this calculation can be reduced to a polynomial of degree smaller than $n$ by exploiting symmetry in the genome space and regularity in the mutational distances between elements of $j \& i$, as detailed in the second line of (3.15):

\(^{10}\)I thank Simon Haines for the derivation of this computational short-cut.
\[
P(|q| = j || p| = i) = \frac{1}{\#j} \left( \sum_{q \in i} \sum_{p \in j} \mu^{\rho(p,q)} (1 - \mu)^{n - \rho(p,q)} \right) \]

\[
= \sum_{k=0}^{n-i} \binom{n-i}{j-i+k} \binom{i}{k} \mu^{j-i+2k}(1-\mu)^{n-j-i+2k}
\]

Figure 3.1 shows mutation dynamics among genome classes. Plot 3.1a visualises M for genomes of length \( n = 4 \), while plot 3.1b shows M for \( n = 8 \). In both cases the rate of mutation is \( \mu = 0.1 \). These genome lengths are much shorter, and the mutation rate much higher, than those that will feature in the model, but demonstrate the dynamics of mutation between classes clearly. For all genome classes mutation is most likely to result in an offspring who belongs to the same class (high probability squares on the counter diagonal of both matrices), but this probability is lower for more extreme genome classes (i.e. the bottom left and top right squares are lower probability than the remaining squares on the counter diagonal), and this asymmetry is stronger where \( n = 8 \) than where \( n = 4 \). Mutation pushes inwards; it degrades information content, and does so with greater force where genomes are larger. This is how the mutation regime imposes costs to specialisation. For example, consider the outcome of a single point mutation. In a strongly biasing genome that contains all 1s, a single mutation necessarily replaces one of those 1s with a 0, resulting in a genome that is less biasing. Conversely, consider a genome that contains only 60% 1s: since mutation is equally likely at any locus, there is a 40% chance that the mutation will switch a 0 into a 1, increasing information content, and only a 60% chance that it will decrease information content by switching a 1 into a 0. Degradational mutation is a natural consequence of additive polygenic inheritance, and is an empirically motivated (e.g. Sanjuán et al., 2004; Eyre-Walker and Keightley, 2007), principled way to capture costs to specialisation.
Chapter 3 The Evolution of Inductive Biases for Culture: A Formal Analysis

Figure 3.1: Mutation dynamics among genome classes (M). Black circular markers indicate the \( ij \)'th coordinate. The colour of the surrounding square gives \( m_{ij} \)

With this we have a mathematically well-behaved regime for representing biological inheritance in the co-evolutionary model, capturing crucial assumptions, and subject to just two significant free parameters: \( n \), the size of a genome, and \( \mu \) the probability of mutation at a given locus.

3.1.3 Cultural Transmission

The representation of cultural transmission in the model is formally fairly simple. I adopt the Iterated Learning model (Kirby et al., 2007; Griffiths and Kalish, 2007), and assume single-teacher vertical cultural transmission: the data observed by a learner at generation \( t \) are generated by a single randomly chosen learner from the previous generation. Though this is a simplifying assumption\(^\text{11}\), it has been widely adopted in similar models (e.g. Griffiths and Kalish, 2007; Kirby et al., 2007; Smith and Kirby, 2008), and has been shown to facilitate conclusions that generalise to more complex transmission dynamics under certain assumptions (Griffiths and Kalish, 2007). Random selection of teachers ensures that new learners receive a representative sample of the linguistic make-up of the previous generation, and allows for the possibility that one teacher can provide data for more than one learner, for example. Transmission dynamics are formalised in the next section (3.1.4), since the evolutionary analysis presented there provides useful notation.

\(^{11}\)I will consider a model that allows learning from multiple teachers (see e.g. Smith, 2009; Burkett and Griffiths, 2010) in the next chapter.
3.1.4 Evolutionary Analysis

Here I present a framework for answering the question: what are the dynamics of an evolutionary system composed of the above components? There are of course many mathematical formalisms and computational techniques for exploring evolutionary dynamics. I make use of the replicator dynamics framework, by adopting a version of the discrete-time replicator-mutator equation to suit the co-evolutionary scenario at hand.

3.1.4.1 Replicator Dynamics

The replicator-mutator equation (Nowak, 2006) is a well-understood (see e.g. Cressman and Tao, 2014), general evolutionary framework: an equivalence between it and the Price equation has been demonstrated, from which various other evolutionary frameworks such as the Lotka-Volterra equation and quasi-species equation can be obtained as special cases (Page and Nowak, 2002), lending confidence to the generality of results obtained in this framework. In broad terms, I wish to track the dynamics over time of two quantities: the genetic makeup and the linguistic behaviour of populations. This requires methods to study the long-term frequency in the population of: each equivalence class $i$ of genomes, and each of the two possible languages $L_0$ & $L_1$. That is, I wish to calculate, at discrete time steps, the distribution of genetic and linguistic variants in an idealised population of language learners.

First, allow the distribution of genome classes at generation $t$ to be given by the vector $g = (g_0, g_1, \ldots, g_n)$, where $g_i$ gives the proportion of the population at generation $t$ made up by genomes that belong to class $i \in G^n/\sim$, and all $g_i$ are probabilities so that $0 \leq g_i \leq 1 \forall i = 0, 1, \ldots, n$, and $\sum(g) = 1$. Similarly, let $x = (x_0, x_1)$ be a probability vector denoting the frequency in the population at time $t$ of the two possible grammars, such that $x_0$ and $x_1$ give the proportion of the population that has induced grammars $L_0$ and $L_1$ respectively. Again $0 \leq x_0, x_1 \leq 1$, and $\sum(x) = 1$. I will use superscripts to denote time points, so that, for example, the genetic distribution at generation $t + 1$ is given by $g^+ = (g_0^+, g_1^+, \ldots, g_n^+)$, and a distribution of languages at generation $t - 1$ is given by $x^- = (x_0^-, x_1^-)$. I wish to numerically obtain measures $g^* = (g_0^*, g_1^*, \ldots, g_n^*)$ and $x^* = (x_0^*, x_1^*)$ that correspond to stable equilibrium distributions in the equations.

---

12My use of superscripts here deviates slightly from common practice. In general such models need only denote two relative time points $t$ and $t + 1$, and so just use the bar superscript to mark $t + 1$. However, here computations require that I refer to generations before and after $t$, so I have modified the notational convention to suit.
that follow. Formally, these measures must satisfy (3.16):

\[ g = g^*, x = x^* \iff g = g^+, x = x^+ . \]  

(3.16)

Though there are various techniques for finding such stable points, I’ll take the following approach: write down equations that determine \( g^+ \) and \( x^+ \) from previous distributions \( g, x, \) and \( x^- \); specify initial distributions \( g^0 \) and \( x^0, \) and; iterate those equations until equation (3.16) is satisfied. The top level recursion, which determines the change in frequency of genome class \( i, \) is given in (3.17):

\[ g^+_i = \frac{1}{\phi} \sum_{j=0}^{n} g_j f_j(x^-, x) m_{ji} . \]  

(3.17)

Of the terms to the right of the sum, two are already defined: \( g_j \) gives the proportion of the population at time \( t \) made up by genomes belonging to class \( j, \) and \( m_{ji} \) denotes the likelihood that during reproduction a genome of class \( j \) will mutate into a genome of class \( i. \) The term \( f_j(x^-, x) \) is new, and represents the fitness or rate of reproduction at time \( t \) of a genome that belongs to class \( j, \) as determined by linguistic distributions \( x^- \) and \( x. \) The quantity \( \phi \) just represents the average fitness of the population, so that:

\[ \phi = \sum_{i=0}^{n} f_i(x^-, x)g_i . \]  

(3.18)

Calculating fitness in this model is arguably the trickiest part, since it is here that the assumption of cultural transmission adds complexity.

3.1.4.2 Calculating Fitness

Here I describe a crucial assumption: the model of cultural evolution I have outlined is used to calculate fitness in the model of biological evolution. A learner’s fitness is dependent upon how well she communicates with the rest of the population; communicative success is dependent upon the language she speaks; the language she speaks is determined by the inferences she makes after observing linguistic data. Crucially, cultural transmission implies a distribution over the linguistic data a learner is likely to encounter. This has knock-on effects at each step of the process, from inference to communication. Interestingly, one way to view the full co-evolutionary model is as a nested adaptive process. Biological evolution can be understood as an inferential process,
gathering information about the environment by adjusting gene frequencies to match the requirements of some adaptive fitness function. Cultural evolution can likewise be understood as an inferential process, gathering information about the population’s cognitive biases by adjusting the distribution of mental representations and utterances. Individual learning can of course also be understood as an inferential process, as I have outlined above. In fact, the formal similarity between the replicator dynamics framework I have adopted, and the model of Bayesian inference I have also adopted, has been noted previously (Harper, 2009). In that sense the nested understanding of this model is direct: we have an instance of a process inside itself; biology being the outer wrapper, and cultural evolution through Bayesian inference being the inner core.

How is the model of cultural evolution used to calculate fitness in the biological model? A learner’s fitness is dependent upon its linguistic behaviour, not its genetic make-up (i.e. its phenotype, not its genotype), and is proportional to the frequency of learners in the current population who have induced the same grammar (proportional to $x_0$ or $x_1$). This assumption reflects the idea that individuals with similar languages communicate well and achieve an advantage; it reflects the adaptationist premise that natural selection for communication has shaped cognition, and instantiates c-induction.

To calculate this quantity, we must obtain for each language the probability that a learner endowed with a genome from class $i$ would induce that language, and use these probabilities to weight a sum of the fitness benefits that would be obtained in either event. As shorthand, allow $a_i^0$ and $a_i^1$ give the overall probability that a learner with a genome from class $i$ will ultimately induce language $L_0$ and $L_1$ respectively. Following similar models (e.g. Nowak et al., 2001), I assume communicative accuracy reflects the probability that any randomly chosen member of the population would share the speaker’s language. The communicative accuracy of an individual who speaks a given language is directly proportional to the frequency of that language in the current population. Thus, fitness is given by:

$$f_i(x^-, x) = a_i^0 x_0 + a_i^1 x_1$$

(3.19)

The outcome of language acquisition is influenced by the data the learner observes, therefore $a_i^0$ and $a_i^1$ must depend upon these data. Because language is culturally transmitted in this model, the data a learner observes depends upon the linguistic make-up of the previous generation ($x^-$) and the dynamics of transmission. Since learners
could end up being taught by any member of the previous generation, and the teacher could produce a range of datasets, and those datasets could lead to induction of either language, there are three domains of uncertainty that must be accounted for when computing acquisition probabilities $a_0$ and $a_1$.

To account for these uncertainties, I must first consider the possible datasets a learner could observe and ask how likely is each given the languages spoken by members of the previous generation, $x^-$. This is made simpler by the fact that production of and inference over a dataset is invariant to the ordering of individual utterances, dependent only on the count $y$ of utterances $S \rightarrow XY$. That means we need only consider the $N + 1$ possible counts $y = 0, 1, \ldots, N$, while also accounting for the number of possible datasets with that count. The probability that a learner at generation $t$ will observe a dataset with count $y$, which I will denote by $P(y|x^-)$, is dependent on the relative frequencies of the two languages in the previous population, and the likelihoods for producing the dataset under each grammar. Since the model of cultural transmission assumes that each learner samples a teacher at random from the previous generation, the probabilities of learning from a speaker of $L_0$ or $L_1$ are given by $x_0^-$ and $x_1^-$. Therefore, the probability of observing a particular count $y$ is:

$$P(y|x^-) = P(y|L_0) \ x_0^- + P(y|L_1) \ x_1^-$$

(3.20)

where the data likelihoods $P(y|L_0)$ and $P(y|L_1)$ are as defined in equations (3.5) and (3.6) respectively. Thus, equation (3.20) defines the distribution over data implied by cultural transmission.

Finally, we must be able to compute acquisition probabilities given data and a genome class. How to compute posterior probabilities given a dataset with count $y$ and an inductive bias $\alpha$ has already been shown in section 3.1.1. These probabilities can be simply re-written to account for assessing a genome class rather than $\alpha$. For clarity:

\footnote{This is because utterances are assumed to be statistically independent and identically distributed, both in production and during inference.}
\[ P(L_0 | y, |q| = i) = \frac{p^y(1 - p)^N - y(n - i)}{p^y(1 - p)^N - y(n - i) + p^{N - y}(1 - p)y_i} \]  \hspace{1cm} (3.21)

\[ P(L_1 | y, |q| = i) = \frac{p^{N - y}(1 - p)^y i}{p^y(1 - p)^N - y(n - i) + p^{N - y}(1 - p)y_i} . \]  \hspace{1cm} (3.22)

For sample learners, these posterior probabilities translate directly into acquisition probabilities, so that
\[ P_{\text{Learn}}(L_1 | y, |q| = i) = P(L_1 | y, |q| = i). \]

For MAP learners:
\[
P_{\text{Learn}}(L_1 | y, |q| = i) = \begin{cases} 
1, & \text{if } P(L_1 | y, |q| = i) > P(L_0 | y, |q| = i) \\
0.5, & \text{if } P(L_1 | y, |q| = i) = P(L_0 | y, |q| = i) \\
0, & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (3.23)

and, in both cases, \[ P_{\text{Learn}}(L_0 | y, |q| = i) = 1 - P_{\text{Learn}}(L_1 | y, |q| = i). \]

With these ingredients we can now formally define \( a_0^i \) and \( a_1^i \), which reflect a sum over all possible datasets, weighted by the distribution implied by cultural transmission, and by the respective acquisition probabilities:

\[ a_1^i = \sum_{y=0}^{N} P(y|x^-)P_{\text{Learn}}(L_1 | y, |q| = i) \]  \hspace{1cm} (3.24)

\[ a_0^i = 1 - a_1^i \]  \hspace{1cm} (3.25)

This completes the calculation of fitness in the model. In summary, to calculate the marginal probabilities that a learner endowed with a genome from class \( i \) will end up speaking each of the possible languages, we must: range over the possible datasets a learner could encounter; calculate for each dataset its likelihood given the linguistic make-up of the previous generation as in (3.20), and; compute the acquisition probabilities for each language given that dataset. Combining these quantities gives probabilities to the outcomes of language acquisition given the dynamics of cultural transmission and the bias of the learner, which in turn allows us to compute the fitness benefits a learner is likely to gain.

For completeness, it is also worth stating the expression for fitness in full as a direct
function of model parameters. Though the formulas that follow appear hairy, it is notable that the fitness calculations can be stated so concisely at all, since a great deal of the model’s complexity resides in this element. The reader is free to skip these expressions and refer instead to the definition above. Generalising over the MAP and sample models initially, fitness can be stated as follows:

\[
fi(x^-, x) = \sum_{y=0}^{N} P(y|x^-) \left( [P\text{\textunderscore learn}(L_0|y) x_0] + [P\text{\textunderscore learn}(L_1|y) x_1] \right). \tag{3.26}
\]

Fitness in the sampling model is given exactly by (3.27):

\[
fi(x^-, x) = \sum_{y=0}^{N} \binom{N}{y} \left( p^y(1-p)^{N-y} x^-_0 + p^{N-y}(1-p)^y x^-_1 \right)
\times \frac{p^y(1-p)^{N-y}(n-i)x_0 + p^{N-y}(1-p)^y x_1}{p^y(1-p)^{N-y}(n-i) + p^{N-y}(1-p)^y} \tag{3.27}
\]

The expression for fitness in the MAP model is a little trickier, since MAP acquisition probabilities amplify posterior probabilities. By using posterior odds rather than probabilities, and noting that \( \frac{a}{b+\infty} = 0 \) for \( a > 0, b > 1 \), fitness for MAP learners can be written as follows:

\[
fi(x^-, x) = \sum_{y=0}^{N} P(y|x^-) \left( \frac{x_0}{1 + O_L^{\infty}} + x_1 \frac{1}{1 + O_L^{\infty}} \right) \tag{3.28}
\]

\[
= \sum_{y=0}^{N} \binom{N}{y} \left( p^y(1-p)^{N-y} x^-_0 + p^{N-y}(1-p)^y x^-_1 \right)
\times \frac{x_0}{1 + \left[ \frac{i-y}{y^2} \right]^{2y-N} + \infty} + \frac{x_1}{1 + \left[ \frac{i+y}{y^2} \right]^{2y-N} + \infty} \right) \tag{3.28}
\]

Almost all calculations required to compute recursion (3.17), which describes cognitive evolution in the model, have now been specified. To complete the model, we must be able to compute the recursion for keeping track of the distribution of languages.
3.1.4.3 Tracking the Linguistic Distribution

I must also be able to determine the population wide distribution of languages at a given time point. Given the assumptions that extant linguistic structure is determined directly by learners and that language is culturally transmitted, this quantity depends upon the predispositions of the learners in the current population ($g$) and the linguistic variation at the previous generation ($x^{-}$). In fact, the ingredients for this quantity have already been seen above: the probability that a learner will induce a particular language given a particular genome class $i$ and a dataset with count $y$ was given in equations (3.21) & (3.22); likewise, the probability that a learner will observe that dataset given the linguistic variation at the previous generation was given in eq. (3.20). These quantities are captured by $a_{i}^{0}$ and $a_{i}^{1}$, so the linguistic distribution can be stated as:

\[
x_{1} = \sum_{i=0}^{n} g_{i} \sum_{y=0}^{N} P(y|x^{-}) P_{learn}(L_{1}|y, |q| = i)
\]

\[
= \sum_{i=0}^{n} g_{i} a_{i}^{1}
\]

(3.29)

\[
x_{0} = 1 - x_{1}
\]

(3.30)

I have now presented the full co-evolutionary model, and a pair of coupled recursions (3.17) and (3.29) for analysing its dynamics. The next section presents results for various analyses of the model.

3.2 Results

3.2.1 Numerical Analysis & Initial Conditions

The model described above defines a system of non-linear recursion equations which capture the dynamics of an evolutionary system whose behaviour I wish to examine. Deriving a closed form solution to that system is impracticable. As such numerical solutions must be obtained: stable points in these equations that represent configurations which, once reached, would endure. Though various techniques can be applied to search for such stable points, I favour a fully dynamic co-evolutionary analysis for all the reasons outlined in the previous chapter. This approach requires that I decide upon
some plausible set of starting conditions, substitute the variables defined above with appropriate numerical values, and allow the system to evolve: iterate recursions (3.17) and (3.29) until (3.16) is satisfied. To the extent that the starting conditions are plausible, the solutions obtained \((g^* \& x^*)\) represent inevitable outcomes of evolution in this system. For each variant of the model examined, I will focus on three sets of starting conditions, each of which reflects slightly different theoretical assumptions about the initial populations and the cognitive capacities under study.

3.2.1.1 Binomial Initial Conditions

First, I will consider initial populations in which each genome class exists at a frequency determined by the mechanics of mutation in this genome model. As shorthand, I will refer to this starting distribution as Binomial. That distribution reflects the prior probability of each genome class \(g_0^i \sim \binom{n}{i}\) as defined by the genome model: were no selective forces operating, mutation would eventually ensure this distribution emerged through drift\(^{14}\). Theoretically, this case might be equated with a population in which the relevant genetic variation exists, but has not heretofore been subject to selection pressures relating to language. The start of the simulation reflects the onset of such selective pressures.

3.2.1.2 Neutral Initial Conditions

The second set of starting conditions represents an initial population in which all learners are entirely neutral, so that \(g_{n/2}^0 = 1\) while \(g_i^0 = 0\) for \(i = 0, 1, \ldots, (n/2) - 1, (n/2) + 1, \ldots, n\). This is arguably the most conservative assumption with respect to the nativist case: it might be interpreted theoretically as a scenario in which the cognitive capacities now under selection were not accessible to the population before the onset of the process. I will refer to this as Neutral conditions.

3.2.1.3 Nativised Initial Conditions

Finally, I will consider a third set of starting conditions in which all individuals are assumed to have a genome that encodes a hard constraint in favour of \(L_1\), so that \(g_n^0 = 1\)

\(^{14}\)This distribution is defined by \(n\) but invariant to \(\mu\), so long as \(\mu > 0\). That is, the frequency of each genome class \(i\) in the stationary distribution associated with the mutation matrix \(M\) is always directly proportional to \(\binom{n}{i}\). Higher mutation rates simply increase the rate of convergence to that distribution.
while $g_i^0 = 0$ for $i = 0, 1, \ldots, n - 1$. Several theoretical interpretations for this set of conditions are possible. For instance, this might reflect the idea of *exaptation*: some biasing cognitive capacity already exists in the population owing to an evolutionary history that did not reflect selective pressures for language. That capacity is recruited by the linguistic system and subsequently maintained by the associated selective pressures\footnote{Note I assume that whatever forces had heretofore held that set of genomes at fixation no longer operate. Such a scenario is consistent with the idea that linguistic communication has replaced an earlier non-linguistic communication system, for instance}. This assumption is intuitively at odds with the evolutionary rationale for nativism, since it presumes the capacities to be explained already exist. However, exaptation is often cited in defence of nativism: for example, Fitch (2011b) argues that "core computational components used today in language could have originally served non-linguistic functions such as motor control, non-verbal thought, or spatial reasoning" (Fitch, 2011b, pp. 1). More importantly, nativised initial conditions will prove to be a useful comparison with the more typical assumptions captured by Binomial and Neutral starting conditions.

In summary: the Binomial starting conditions reflect the pressures of the genome model and the inbuilt preference for neutrality; Neutral assumes a totally unbiased initial population, and; Nativised assumes the population is already converged on the most extreme genome class. Before presenting results for the full MAP & sample learner models, I consider two important baseline analyses that will offer vital comparisons with the full model later. The first (Biology Only - section 3.2.2.1) removes language from the model entirely, and implements selection directly on the genotype, as a way to confirm the suitability of the genome model for evolutionary analysis. The second (No Learning, No Culture - section 3.2.2.2) removes learning and transmission from the model, as a way to isolate the consequences of cultural learning in the full model.

### 3.2.2 Two Baseline Analyses

#### 3.2.2.1 Biology Only

Results in this section concern a stripped down version of the model presented above. Here language learning and cultural transmission are removed from the model entirely, and selection acts directly on the genotype of the individual. Given the assumption that
stronger biases should be \textit{a priori} less probable than weaker biases, a natural concern is that the genome model systematically prohibits the evolution of strongly biasing genomes, especially where \( n \) is larger, because they are simply too improbable, or difficult to maintain against degradational mutation. To allay this concern, Figure 3.2 shows the results of a model in which genomes that encode the strongest priors are explicitly selected for. In this model the fitness, \( f_i \), of a genome belonging to class \( i \) is given by:

\[
f_i = i .
\]  

(3.31)

This selects directly for genomes with more 1s. Selecting for 1s rather than 0s is an arbitrary choice, owing to the symmetry of the genome space. Since selection is not frequency dependent, the recursion for genetic types becomes:

\[
g_i^+ = \frac{1}{\phi} \sum_{j=0}^{n} g_j f_j m_{ji},
\]  

(3.32)

where \( \phi = \sum_{j=0}^{n} i \times g_i \) is the average fitness of the population, and the condition for system-wide stability becomes:

\[
g = g^* \iff g = g^+.
\]  

(3.33)

Shown in the left hand side panels of figure 3.2 in red are the initial \( (g^0) \) distributions of genome classes assumed. The topmost panel shows Binomial initial conditions, the middle panel Neutral, and the bottom panel Nativised. Shown in blue is the stable distribution \( (g^*) \) of genomes associated with each of these starting conditions once the model has converged: the distribution given by blue bars is a fixed point in this evolutionary system.
Figure 3.2: Evolution in the Biology Only model. Each row shows initial (red) and stable (blue) distribution of genome classes in the left hand column, and the change over time of the mean (and one standard deviation) genetic make-up of the population in the right hand column. In all cases extreme genomes come to dominate the population and remain stable thereafter. $n = 100$, $\mu = 0.001$

With these parameters: the Binomial model converged after 3598 iterations, the Neutral model converged after 3762 iterations, and the Nativised model converged after 3349 iterations. All starting conditions lead to the same stable distribution shown in blue. That distribution strongly favours the most extreme genomes. Letting $\Phi^* = \sum_{i=0}^{n} (g_i^* \times i)$ be the average number of 1s in the genomes of the population, $\Phi^* = 91.37$ in all three cases.

The right hand side columns of figure 3.2 show the associated time course plots. So, for instance, the top right panel shows how $\Phi$ evolves over time given Binomial initial conditions, while the right hand side panel in the second row shows the same plot for
Neutral initial conditions. Evolution has no problem stabilising the extreme genomes in the population: they come to dominate gradually but swiftly. This demonstrates that, where there is selection that favours the extreme genomes, they can be maintained against mutation pressure.

3.2.2.2 No Learning, No Transmission

Here cultural learning is removed from the main model, but the linguistic phenotype is kept. Each learner acquires its grammar probabilistically on the basis of its prior preferences, determined by its genome, irrespective of the languages acquired by other members of the population. This scenario might be compared to the evolution of an innate communication system, for instance: it reflects the basic adaptationist evolutionary model for the origins of innate traits that don’t involve cultural learning. Selection favours coordination rather than any particular language: this ensures the case is as closely comparable to the cultural model as possible, and implements frequency-dependent selection. It is essentially the scenario envisaged by the traditional adaptationist account of the evolution of language related cognition (e.g. Pinker and Bloom, 1990) that assume the linguistic phenotype grows like any other organ rather than being inherited culturally under biased learning. These assumptions simplify the mathematics somewhat. The recursion for language frequencies is no longer a recursion, but becomes:

\[ x_1 = \frac{1}{n} \sum_{i=0}^{n} g_i \times i , \]  

and \( x_0 = 1 - x_1 \). Acquisition probabilities simplify too, so that \( P_{\text{learn}}(L_1) = \frac{i}{n} \) and \( P_{\text{learn}}(L_0) = \frac{n-i}{n} \). Fitness is determined by communicative success; by the probability that a learner with a genome from a given class would acquire each language, weighted by that language’s respective frequency in the population, so that:

\[ f_i(x) = \frac{i}{n} x_1 + \frac{n-i}{n} x_0 . \]
With these assumptions the genetic recursion is:

\[
g_i^+ = \frac{1}{\phi} \sum_{j=0}^{n} g_j f_j(x) m_{ji},
\]

(3.36)

with \( \phi = \sum_{j=0}^{n} f_j(x) r_j \) again being the average fitness of the population. Again, system wide stability depends only on the population’s genetics, so that

\[
g = g^* \iff g = g^+.\]

(3.37)

and the stable distribution of languages is:

\[
x_1^* = \frac{1}{n} \sum_{i=0}^{n} g_i^* \times i,
\]

(3.38)

\[
x_0^* = 1 - x_1^*.\]

(3.39)

Figure 3.3 shows evolution in this model. Here again extreme genomes are fixed in the population. All three sets of starting conditions lead to the same stable distribution, with \( \Phi^* = 90.47 \) and \( x_1^* = 0.9 \), which reflects a strong innate constraint favouring language \( L_1 \), and an overwhelming majority of language \( L_1 \) users.

Interestingly, this model takes slightly longer to converge than did the Biology Only model. Binomial converged at generation 5176; Neutral at 5096, and; Nativised at 5096. The model suggests that evolution in a coordination problem with no cultural learning can lead to strong innate constraints on cognition. This result aligns with several reviewed in the previous chapter, such as those reported by Christiansen et al. (2011) and by Lachlan and Feldman (2003), and is in line with the wide distribution of innate communication systems among non-human animals. This model extends previous models by allowing a broad range of graded genetic types, and degradational mutation between these. Even under these conditions, evolution supports nativism for a-cultural communication systems.
Chapter 3  The Evolution of Inductive Biases for Culture: A Formal Analysis

Figure 3.3: Evolution in the No Learning No Culture model. Each row shows initial (red) and stable (blue) distribution of genome classes in the left hand column, and the change over time of the mean (and one standard deviation) genetic make-up of the population in the right hand column. In all cases extreme genomes come to dominate the population and remain stable thereafter. \( n = 100, \mu = 0.001 \)

With these baseline analyses in hand, I now turn to the full co-evolutionary analysis, first exploring evolution in populations of sample learners, then in populations of MAP learners.

3.2.3 Full Co-evolutionary Analyses

Results here concern the main co-evolutionary model outlined above. Given the centrality of these models to my aims, a more thorough exploration of their behaviour over the range of possible parameter values is desirable. Exhaustive analysis is impractical, and would also involve much redundancy. Cultural transmission introduces two new parameters: \( N \), the number of utterances each learner observes, or the bottleneck on transmission, and \( p \), the degree to which utterances are diagnostic of the language which generated them. Together these parameters define how rich are the data learners
observe: more data (higher \( N \)) and more diagnostic utterances (higher \( p \)) provide the learner with a more accurate signal of the generating language. The parameters of the biological model determine the relative strength of degradational mutation pressure, and as such the difficulty of the evolutionary task: longer genomes (higher \( n \)) and more mutation (higher mutation rates \( \mu \)) penalise stronger biases more{superscript 17}. In what follows, I will vary \( N \) and \( n \), while holding \( p = 0.95 \) and \( \mu = 0.001 \) at constant values, and begin with the default{superscript 18} case: lower quality data (\( N = 2 \)) (this instantiates the impoverished stimulus central to nativist reasoning) and a higher costs to strong biases (\( n = 100 \)). Thereafter I will explore different parameter values as and when necessary.

### 3.2.3.1 Sample Learners

Figure 3.4 shows results for the sample learner model where learners observe sparse (\( N = 2 \)) data{superscript 19}. To break the symmetry of the model, \( x^0_0 = 0.55 \) represents an initial frequency advantage to language \( L_1 \), of a degree that might be expected through stochastic fluctuations in finite populations. Given the symmetry of the model, it is sufficient to explore this case and not the reverse (e.g. \( x^0_1 = .45 \)) since results are qualitatively identical. Here, the mean prior bias in the population is given by:

\[
\Phi = \frac{1}{n + 1} \sum_{i=0}^{n} g_i \times i .
\]  

(3.40)

For example, \( \Phi = 1 \) would represent a population converged on a hard constraint that rules out language \( L_0 \), while \( \Phi = 0.5 \) represents a population that is neutral overall. Values of \( \Phi \) intermediate between these represent populations that share a weaker innate inductive bias.

---

{superscript 17}Since mutation degrades information content, more mutation degrades information (bias content) faster. Technically, higher mutation rates push the distribution of genome classes towards its stationary state (which favours neutral genomes) at a greater rate.

{superscript 18}This scenario is default by stipulation, though reflects conservative assumptions.

{superscript 19}Initial distributions of genome classes are not plotted in this figure (and subsequent equivalent figures): for binomial initial conditions, \( g^0 \) would sit right on top of \( g^* \); for neutral and nativised initial conditions, visualising \( g^0 \) would change the scale of the plot such that \( g^* \) is not clear.
Figure 3.4: Evolution in the sample learner model. **Left hand column panels:** initial genome class distributions ($g^0$) are not shown, since the final distributions ($g^*$) would be obscured. As before, the top row corresponds to Binomial initial conditions, the middle Neutral, and the bottom Nativised. **Right hand column panels:** the green line shows the mean prior bias ($\Phi$) in the population; one standard deviation in the distribution of biases in the population is shown above and below the green line by the shaded area; the purple line shows $x_1$, the population frequency of language $L_1$. $n = 100$, $\mu = 0.001$, $N = 2$, $p = .95$
Two patterns of results are clear from these figures. First, results for the Binomial and Neutral conditions are essentially identical. Both converge on the same stable distribution of genomes, at roughly similar speeds. That distribution reflects the dynamics of mutation in the model, such that $g^* = g^{\text{Binomial}}$, or rather, $g_i^* = \frac{1}{2^{n_i}} \times \left(\begin{array}{c} n \\ i \end{array}\right)$. On average the stable population is not biased towards either language, so that $\Phi^* = 0.5$.

The cultural distribution closely follows the genetic distribution: the purple line showing $x_1$ closely tracks the green line in all cases. In sum, no genetic or cultural universals emerge. Selection for improved communication under these conditions does not lead to the emergence of strong domain-specific cognitive constraints. It is unable to overcome degradational mutation pressure, and therefore unable to fix coordinated biases in the population. Under exactly the same conditions, evolution in the No Learning No Culture model did fix strong cognitive constraints in the population. Cultural transmission through expression and induction prevents this outcome. Why?

In general, the selective value of a cognitive bias is, for the sample learner, proportional to its strength. Inductive biases are fitness-enhancing when they increase the probability of acquiring the common linguistic variant. This probability, and therefore the fitness advantage, is proportional to the strength of the inductive bias. This means that weak biases, which are most likely to emerge through (and least vulnerable to) mutation, afford equivalently weak fitness advantages: the biased sample learner accrues proportional gains. Selective differentials are stronger when the cultural distribution is more skewed towards one language: it is more valuable to be strongly biased towards learning a given language if that language is spoken by everybody else in the population. Without the amplification effect of culture, linguistic universals only emerge when everybody possesses strong biases. Unless they are assumed to pre-exist, strong biases must evolve from weaker biases. However, weaker biases don’t lead to a cultural universal. For this reason, their fitness benefits do not outweigh the costs imposed by mutation for long enough to scaffold the evolution of stronger biases.

However, individuals in the No Learning No Culture model also acquire traits with probability proportional to their cognitive bias, and thus also accrue proportional gains, yet this does not prevent nativisation from initial neutrality. Why then is this impossible in the initially neutral sample learner model? Some property of vertical cultural transmission through sample learning has, from the outset, put the brakes on the co-evolutionary spiral that eventually leads to nativisation.
Consider the beginnings of this co-evolutionary process in detail. From initial populations that are genetically and linguistically neutral or near-neutral, selection for communicative coordination works to increase the frequency of genome classes that encode stronger biases (in this case toward language $L_1$) and thus reduce intra-population linguistic variation. Mutation dynamics work in precisely the opposite direction: they act to degrade the average information content in the population’s genomes, or to decrease the frequency of strongly biasing genome classes. To overcome mutation dynamics, selective differentials must be sufficiently skewed (in favour of biasing genomes): relative to less biasing genomes, more biasing genomes must enjoy sufficiently greater fitness benefits. Assuming language $L_1$ is more common than $L_0$, fitness differentials are determined by the relative likelihood that an individual with a given genome will end up using language $L_1$. In the No Learning No Culture model, acquisition probabilities depend only upon the learners’ predispositions. Skewed fitness differentials are guaranteed, since predispositions vary. In the Sample learner model, acquisition probabilities depend upon predispositions and the data learners observe. Crucially, whatever influence data have on the outcome of acquisition is common to all learners. Therefore fitness differentials in the sample learner model may be weaker than in the No Learning No Culture model. When learning from a previous generation that spoke both languages in roughly equal proportion, the biased sample learner is less likely to acquire $L_1$ than is the equivalently biased individual in the No Learning No Culture model. The biased sample learner’s fitness benefit is diluted by the influence of linguistic variation at the previous generation. I will refer to this property of co-evolution from neutrality without the amplification effect as the sub-proportional gains (sub-PG) problem. The historical state of the linguistic system introduces a lag in the genetic system’s response to selective pressures. On an evolutionary trajectory towards less linguistic variation, the co-evolutionary process that includes culture, but not the a-cultural equivalent, faces the sub-PG problem.

The second pattern of results for sample learners is shown in the bottom panels of figure 3.4. Here, the stable distribution of genome classes is concentrated on those that encode the strongest biases towards language $L_1$. In this case the mean bias in the population converges to $\Phi^* = 87.28$, a strong preference for language $L_1$, and that language is near universal ($x_1^* = 0.89$). Note that $\Phi^* < \Phi^0$: while selection has been able to maintain strong priors in the population, the hard constraint $\alpha = 1.0$ enjoyed by the initial
population has been degraded into a strong bias, and the associated linguistic distribution is near universal, but not completely universal. The degree of degradation (and therefore the ultimate strength of bias and statistical strength of the associated linguistic universal) to be expected is determined by a trade-off between selection strength and mutation pressure: I will return to this point shortly. Crucially, though, the difference between the first and second pattern of results (caused by a difference in the genetic make-up of the initial populations) in the sample learner model is of great significance.

The second pattern of results shows that shared, coordinated strong cognitive biases are a stable configuration of this evolutionary system: where strong biases already exist in the population, they give rise to the linguistic (near) universal which creates a selection pressure that can maintain those biases. The time-course plot for the Nativised model in figure 3.4 shows that this is true: the linguistic universal is caused by the initial genetic make-up of the population, which then plays out a complex dynamic reflecting the balance of selection and mutation pressures, but eventually settles on strong cognitive biases: nativism about the learners in this population is warranted. However, results from the Binomial and Neutral models show that this configuration, though stable, cannot be reached: while they can be maintained, strong cognitive biases in this model are not evolvable.

Figure 3.5 shows results in the sampler model where learners see double the amount of data (N = 4). Results in the Binomial and Neutral conditions are exactly as in the model with half the data (N = 2, figure 3.4). Results for the Nativised condition are qualitatively similar, but quantitatively different: more data allows higher fidelity language acquisition and therefore better coordination. Evolution in this case stabilises slightly weaker priors in the population (Φ∗ = 0.79), which nevertheless still give rise to a linguistic distribution that is skewed as strongly in favour of language L1, with x1∗ = 0.89. Why is the linguistic distribution still so strongly skewed relative to the population’s prior bias, given that Griffiths and Kalish (2007) proved that cultural transmission among sample learners converges to the learners’ prior bias?
Figure 3.5: Evolution in the sample learner model. \( n = 100, \mu = 0.001, N = 4, p = 0.95 \)

Recall that their proof concerned the long-term dynamics of cultural transmission: culture converges to faithfully reflect cognitive biases over \textit{infinitely many} generations. However, larger bottlenecks (more data) attenuate the rate of convergence, so that in a given finite period of time, the linguistic distribution may not be isomorphic to the distribution of cognitive biases in the population. Over time these differences iron out through gradual cultural evolution. However, larger bottlenecks slow down the cultural evolutionary process: where learners have sufficient data to ensure highly faithful transmission, cultural change may be very slow indeed: slow enough to approximate stability in practical terms. For instance, figure 3.6 shows results in the sampler model with the bottleneck doubled again, so that \( N = 8 \).
Chapter 3: The Evolution of Inductive Biases for Culture: A Formal Analysis

Figure 3.6: Evolution in the sample learner model. $n = 100, \mu = 0.001, N = 8, p = .95$

Given the small space of possible languages, this bottleneck is sufficiently high to ensure reliable language acquisition and high fidelity transmission. Results in the Binomial and Neutral conditions don’t change: both converge to the Binomial distribution: Binomial does so in 60020 generations; Neutral converges to stability after 62489 generations. The picture in the Nativised condition is somewhat more complex. Mutation pressure degrades pre-existing genetic biases as before, but to a larger extent. However, the linguistic distribution, thanks to the higher bottleneck, changes very slowly and continues to reflect the conditions of the initial populations: while pre-existing biases are degraded through mutation, the linguistic near universal brought about by those biases pre-degradation persists. Initially, the pre-existing hard constraints enforced a linguistic universal. Since high fidelity cultural transmission ensures coordination regardless (or near regardless) of cognitive bias, those hard constraints lose some of their selective value, and are degraded.

The distributions shown for the Nativised condition in figure 3.6 are not stable:
Table 3.1: Comparison of Evolved Biases in the Sample Learner Model.

<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Sparse Data ((N = 2))</th>
<th>Rich Data ((N = 8))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hard Evo. ((n = 100))</td>
<td>Easy Evo. ((n = 50))</td>
</tr>
<tr>
<td>Binomial</td>
<td>Unbiased ((\Phi^* = 0.5))</td>
<td>Strong Bias ((\Phi^* = 0.91))</td>
</tr>
<tr>
<td>Neutral</td>
<td>Unbiased ((\Phi^* = 0.5))</td>
<td>Strong Bias ((\Phi^* = 0.91))</td>
</tr>
<tr>
<td>Nativised</td>
<td>Strong Bias ((\Phi^* = 0.87))</td>
<td>Strong Bias ((\Phi^* = 0.91))</td>
</tr>
</tbody>
</table>

after 500,000 generations (not plotted) the model does not converge to absolute stability. However, the degree of continued change is so small as to be negligible in real terms (had I restricted the resolution of \(g_i\) to five 5DP, for instance, the model would have converged after around 60,000 generations). In practical terms, evolution favours weak biases in this scenario \(\Phi^{500000} = 0.54\), for instance). Despite weak prior preferences, the linguistic distribution, which reflects a combination of historical conditions and near-stable cognitive preferences, is concentrated on language \(L_1\), such that \(x_1^{500000} = 0.893\). Given more time these biases might be entirely eroded. However, given the assumption that weaker biases face weaker degradational pressures and therefore will be eroded more slowly, we might assume that in this scenario, on realistic time-scales, the weakest biases are the most probable outcome of cognitive evolution. Again, even in a scenario where cognitive adaptation leads to weak biases, those biases can be maintained when they result from pre-existing biases, but are not evolvable from neutrality. Table 3.1 shows results for the sample learner model across the parameter dimensions I discussed at the beginning of this section.

The pattern of results implied by this table may seem confusing, but is in fact reasonably straightforward. Arguably the most important case is the default scenario, in which learners observe impoverished data \((N = 2)\), specialisation incurs greater costs \((n = 100)\), and the genetic composition of initial populations is determined by the genome model rather than any additional assumptions (Binomial starting conditions). This scenario engages with nativist reasoning on its own terms (by allowing...
that observable linguistic data are sparse), but makes conservative assumptions about how difficult is the evolution of specialised cognitive systems. Evolution in the Biology Only and No Culture No Learning models under these conditions led to clear and strong nativisation, but does not once cultural transmission is added to the model. With culture, these conditions lead to unbiased learners and no linguistic universals. Something about the process of cultural evolution has intervened in the formation of innate knowledge, such that evolutionary predictions which may be sound in the case of a-cultural communication systems must be tempered in this case.

The evolutionary defence of linguistic nativism only plays out successfully where costs to specialisation are assumed weak, or pre-existing cognitive biases are assumed. Also note that, where it occurs, adaptation in this model takes much longer than adaptation in the equivalent No Learning No Culture model. For example, in the $n = 50$, $N = 2$ case, which is the most favourable to nativisation, cognitive adaptation among sample learners takes roughly an order of magnitude longer (convergence to stability from Binomial initial conditions at 34379 generations) than in the equivalent No Culture No Learning model (convergence to stability from Binomial initial conditions after 2360 generations), and only occurs after an extremely long period of slow change, during which the population is not supporting a linguistic universal. Figure 3.7 shows evolution under these parameters: change is initially so slow because the fitness advantages gained by biased learners are so weak in the absence of a cultural universal. Cultural variation is characteristic of relatively unbiased sample learner populations: even where it is possible, it is extremely difficult for them to escape neutrality.

In some cases evolution stabilises weak cognitive biases: this occurs where the cultural process ensures high fidelity acquisition and transmission, and some biases in favour of the eventually universal language already exist in the population. If costs are assumed low, initially low frequency biases can bring about this effect. If costs are assumed high, the population must be assumed initially subject to a hard constraint.
In summary, analysis of the sample learner model demonstrates that its evolutionary dynamics do not map directly onto standard reasoning about the evolution of cognitive specialisation for a-cultural traits: co-evolution in this model under these conditions can only lead to strong innate constraints if the population of interest is already biased in that direction. In the absence of the amplification effect, culture imposes the sub-PG problem, which prevents co-evolution from neutrality getting off the ground.

### 3.2.3.2 MAP Learners

In this section I present results for the MAP model. Recall the difference between MAP and sample learners: where sample learners induce a language on the basis of its posterior probability, MAP learners compute the same posterior probabilities but induce whichever language maximises this quantity. This is the only difference between the models presented here and in the previous section. Nevertheless, this minor difference in the individual learning process gives rise to an entirely different pattern of evolu-
tionary results, because it gives rise to the amplification effect. Figure 3.8 is comparable to those presented for the sample learner model, and shows results for the default scenario.

![Graph](image)

Figure 3.8: Evolution in the MAP learner model. $n = 100, \mu = 0.001, N = 2, p = .95$

All initial conditions lead to qualitatively similar results: the fixation of weak cognitive biases in favour of, and near universal adoption of, language $L_1$. Binomial and Neutral conditions lead to the same stable distribution. In the top row, I have included the initial Binomial distribution again in red to demonstrate that, although they look similar, the initial and final distributions are crucially distinct. The stable distribution of genomes is centred on the weakest possible biases towards $L_1$, such that $\Phi^* = 0.55$. Note also that this distribution is right-skewed, favouring biases towards $L_1$ over neutrality. These weak cognitive biases emerge extremely rapidly: the population settles on shared, coordinated weak biases within a few hundred generations. Relative to
evolution under identical conditions in the Biology Only and No Learning No Culture models, the adaptive outcome in the MAP model emerges roughly an order of magnitude faster: culture has sped-up cognitive adaptation. Crucially, despite their strength, these biases support a linguistic universal: language $L_1$ is driven to near-fixation\(^{20}\) in around a hundred generations, and is stable soon after with $x_1^* = .96$. What can explain this pattern of rapid adaptation through fixation of the weakest biases?

Recall Griffiths and Kalish’s (2007) finding that cultural transmission in populations of MAP learners can lead to learners’ cognitive biases being amplified: weak biases can give rise to strongly skewed cultural distributions. Because language $L_1$ is very slightly more common than $L_0$ in initial populations, learners who are more likely to acquire $L_1$ through cognitive bias gain a slight fitness advantage. As the genetic distribution is tipped in favour of genomes that bias learners towards $L_1$, that language very quickly takes over the population because of the amplification effect: on average the population has a slight bias towards $L_1$, which is amplified and gives rise to a linguistic universal. This universal in turn imposes a strong selection pressure for reliable acquisition of $L_1$. In fact, this pressure is as strong as is possible in the model: since $L_1$ is universal, acquiring $L_0$ is a terrible mistake. Because language is transmitted culturally, the data learners then observe are likely to portray $L_1$ over $L_0$, however minimally. Even a weak bias towards $L_1$ then guarantees its acquisition and the associated fitness benefits: the benefits of weak biases are unmasked by the cultural process (I will return to this idea in the discussion section of this chapter).

Since any bias toward $L_1$, weak or strong, guarantees its acquisition and fixation among the population, the population settles on those that satisfy a delicate balance: they must i) minimise the degradational pressure imposed by mutation, whilst also ii) being sufficiently distant from neutral that mutant offspring are unlikely to be neutral or even biased in the wrong direction. Together these pressures favour the weakest biases that are also outside a mutational buffer or danger zone\(^{21}\) around the neutral genomes.

---

\(^{20}\)Here and throughout I’ll take strong statistical tendencies to be universals of language. In the typological literature a distinction is made between absolute and statistical universals. My use of the term should be understood in the latter sense, in that, for example, I’ll discuss values of $x_1 \approx 0.96$ as linguistic universals, though note that some would object to the use of this term when there are any exceptions to the generalisation.

\(^{21}\)It pays to be sufficiently distant from neutral that mutated offspring don’t fall on the wrong side of $\alpha = 0.5$. Technically, this is known as selection for mutational robustness. See e.g. van Nimwegen et al. (1999).
Are the initial conditions building in this result? Both Binomial and Neutral initial conditions privilege genomes that encode neutrality or biases near $\alpha = 0.5$. While this is true, and in my view a sensible default assumption, the outcome isn’t dependent on this initial state. The stable distribution associated with Nativised initial conditions, pictured in the bottom left panel of figure 3.8, is almost identical: it too has mean bias $\Phi^* = 0.55$, and supports a linguistic universal ($x^* = .96$). The associated time-course plot shows an evolutionary pattern very similar to that observed among populations of sample learners with access to rich unambiguous data (figure 3.6): initial hard constraints enforce a linguistic universal and are then degraded into weak biases by mutation pressure, but the universal remains thanks to the cultural process. Again this evolutionary outcome happens quickly: only a thousand or so generations are required for the population to settle on weak biases. These dynamics are reminiscent of, but importantly different from, Boyd and Richerson’s (1985) result concerning the evolution of direct bias. Crucially, the model I have presented allows a continuum of biases to exist in the population, and, as has been repeatedly demonstrated in models of cultural transmission through expression, inference, and induction (e.g. Griffiths and Kalish, 2007; Kirby et al., 2007), skewed cultural distributions cannot be maintained in the absence of bias. These two factors explain why weak biases are stabilised in this case but not theirs. The stabilisation of weak biases in the MAP model, which gives rise to and supports a linguistic universal, is an extremely robust dynamic. Because learners’ biases are amplified over cultural transmission, and weak biases face least cost, they are rapidly stabilised whatever the difficulty of the evolutionary or inferential task.

Figure 3.9 shows evolution in the MAP model where costs to specialisation are relatively low ($n = 50$) and learning is relatively inaccurate ($N = 2$). This is the case most favourable to nativism. Where learners observe little data, the influence of the prior is strong. Where stronger biases require the coordination of fewer genes, they face less destructive mutation pressure. Even with these assumptions, and even when populations are initialised with hard constraints favouring $L_1$, the same evolutionary dynamic is recovered: stabilisation and amplification of defeasible cognitive biases, not strong biases or hard constraints.
Table 3.2 shows outcomes of cognitive evolution in the MAP model over a range of parameters. In no cases are strong innate constraints fixed among the population. In most cases the population quickly settles on weak biases, which invariably support a linguistic universal ($x_i^* \approx 0.95$ in all scenarios). Differences in the strength of the stable biases are clear but minor. The pattern is: stronger costs to specialisation (higher $n$), and less sparse data (higher $N$) each decrease the strength of the resultant cognitive biases.
Table 3.2: Comparison of Evolved Biases in the MAP model.

<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Sparse Data ((N = 2))</th>
<th>Rich Data ((N = 8))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hard Evo. ((n = 100))</td>
<td>Easy Evo. ((n = 50))</td>
</tr>
<tr>
<td>Binomial</td>
<td>Weak Bias ((\Phi^* = 0.55))</td>
<td>Weak Bias ((\Phi^* = 0.58))</td>
</tr>
<tr>
<td>Neutral</td>
<td>Weak Bias ((\Phi^* = 0.55))</td>
<td>Weak Bias ((\Phi^* = 0.58))</td>
</tr>
<tr>
<td>Nativised</td>
<td>Weak Bias ((\Phi^* = 0.55))</td>
<td>Weak Bias ((\Phi^* = 0.58))</td>
</tr>
</tbody>
</table>

3.2.4 Stochastic Simulations in Finite Populations

The analyses above assume idealised populations, and are based on calculations at the mean field level. While the findings of those analyses are precise, they do not account for any factors that might result from the finite nature of human populations. For instance, stochastic effects might overcome or exacerbate dynamics that result from low probability events, or delicate balances between pressures, such as the sub-PG problem in populations of sample learners. Previous co-evolutionary modelling has demonstrated that infinite population models and stochastic simulations in finite populations can lead to results that differ both qualitatively and quantitatively (Komarova and Nowak, 2003). Here I re-implement important cases of the full co-evolutionary models as stochastic simulations in finite populations of \(N\) learners per generation.

The computational details are as follows. The first generation’s genomes are sampled randomly according to their Binomial probability (Binomial initial conditions) under the genome model, and those learners each sample a language according to their prior preferences. Each learner’s fitness is proportional to the number of learners in the population who have induced the same language as that learner. A new generation of \(N\) learners is produced by roulette wheel sampling according to the older generation’s fitness scores and subject to mutation. Each new learner randomly samples a teacher from the previous generation, and learns a language from data produced by that teacher. The older generation is then removed from the simulation. This process repeats for a pre-specified number of generations. Figure 3.10 shows the results of 100
such simulations under default parameters \((n = 100, \mu = 0.001, N = 2, p = 0.95)\). Each point shows the proportion \(x_1\) of language \(L_1\) users in the population, and the mean bias \(\Phi = \frac{1}{N} \sum_{k=1}^{N} \alpha_k\), where \(\alpha_k\) is the prior of the \(k\)th learner, after 1000 generations for MAP (purple markers) and sample (yellow markers) learners.

![Graph showing evolutionary outcomes](image)

**Figure 3.10:** Evolutionary outcomes of stochastic simulations of the mathematical model in finite populations of learners. Markers show the average cognitive bias \(\Phi\) and the proportion \(x_1\) of language \(L_1\) speakers at generation 1000 for sample (yellow markers) and MAP (purple markers) populations. \(N = 100, n = 100, N = 2, \mu = 0.001\)

Results showing strong nativisation and associated linguistic universals would appear in the top right or bottom left corners of this plot, but never occur. Results replicate those of the numerical analyses: populations of MAP learners evolve shared weak innate biases that drive the favoured language to dominance or near-dominance among the population; sample learner populations vary randomly around neutral, evolving weak cognitive biases but no corresponding linguistic universal.

A particularly interesting case concerns evolution among sample learners where costs to specialisation are lower \((n = 50)\). Numerical analysis of the mathematical
Chapter 3  The Evolution of Inductive Biases for Culture: A Formal Analysis

model suggests the outcome of cognitive evolution in these conditions is strong innate constraints, but only after many generations of slow gradual change: the selective value of stronger constraints eventually overpowers mutation pressure and overcomes the sub-PG problem, but escaping initial neutrality is difficult. How does this fine balance play out in finite populations? Figure 3.11 shows the outcome of evolution in these populations (and MAP learner populations under the same conditions) after 5000 generations. In finite populations that small advantage is not sufficient to overpower mutation: whatever low probability events were captured in the infinite population model do not occur in the stochastic simulations of finite populations. Neither sample learners nor MAP learners end up with strong innate constraints.

Figure 3.11: Evolutionary outcomes of stochastic simulations of the mathematical model in finite populations of learners. Markers show the average cognitive bias $\Phi$ and the proportion $x_1$ of language $L_1$ speakers at generation 5000 for sample (yellow markers) and MAP (purple markers) populations. $N = 100, n = 50, N = 2, \mu = .001$
3.3 Discussion

What do the preceding analyses say about the evolution of cognitive faculties specialised for language? The baseline analyses of a-cultural systems I considered bear out evolutionary predictions: selection for improved communication leads to the evolution of strong domain-specific innate cognitive biases that constrain which traits can be acquired. The adaptationist evolutionary rationale for nativism about these behaviours is valid.

I have analysed two models which include cultural transmission: one which includes the amplification effect often associated with cultural transmission (the MAP learner model), and one which does not (the sample learner model). In both cases, the addition of cultural transmission through expression, inference, and induction can radically change the outcome of cognitive evolution. Intriguingly, learning from others intervenes in the formation of innate knowledge in quite different ways according to the assumptions made about the nature of individual learning and the cultural dynamics thus induced. In the sample learner model, under the most plausible conditions, cultural transmission brings the co-evolutionary process to a grinding halt. In the MAP model, culture works wonders, facilitating adaptive short cuts that make strong innate constraints redundant. I’ll discuss each of these evolutionary dynamics in turn, and conclude that on the whole, evolutionary reasoning does not support linguistic nativism as traditionally conceived.

3.3.1 Evolution, Nativism, & Culture without Amplification

Cultural transmission among sample learners does not give rise to the amplification of prior preferences often associated with cultural transmission: rather, the distribution of linguistic forms within a population eventually comes to faithfully mirror learners’ predispositions concerning those forms (Griffiths and Kalish, 2007). As a result of this dynamic, no cognitive adaptation or linguistic universals tend to result from co-evolution.

3.3.1.1 Culture & the sub-PG Problem

Cultural transmission creates a link between linguistic distributions over time. For natural selection to build coordinated strong innate constraints, it must be able to see sufficient phenotypic differences between genetic types: biased learners must be sufficiently
more likely than unbiased learners to acquire common linguistic variants. At least, the biased learner’s fitness advantage (gained through increased likelihood of acquiring the common variant) must outweigh whatever costs are incurred through specialisation. In a co-evolutionary scenario that includes culture but not amplification, the biased mutant offspring of a lesser biased parent population suffers the fitness consequences of linguistic variation among her parents: she learns from a linguistic distribution that, in the absence of amplification, reflects the lesser biased population. I have termed this hypothesis the *sub-proportional gains* problem. To achieve a given improvement in the linguistic distribution, natural selection must effect a *larger* improvement in the genetic distribution.

The *sub-PG* problem has an intuitive interpretation. For example: in well mixed populations that learn culturally, individuals who acquire an uncommon linguistic variant might not have many children of their own, but may still speak to, and therefore influence, children of the more reproductively successful learners. If cognitive biases and cultural experience both contribute to the linguistic phenotype, and biases are not amplified, there is always the possibility that the cultural learner’s biases, despite being the result of positive selection, will be attenuated by experience with the older, pre-selection population.

Whether or not the *sub-PG* problem can be overcome will depend upon the relative magnitudes of selection and mutation costs, and the details of cultural transmission. However, the models I have analysed suggest that: while evolutionary reasoning does in general support nativism about an a-cultural behaviour, it does not support nativism about culturally transmitted behaviour unless there is independent reason to believe the *sub-PG* problem can be overcome. Evolution of strong innate constraints is not a robust outcome of the process, but rather is contingent upon favourable assumptions over and above those required for a-cultural behaviours. Even granting liberal assumptions that allow the evolution of strong innate constraints in this model, there are still further reasons to doubt the credibility of this dynamic as an explanation for the origins of language-specific cognition.

### 3.3.1.2 Further Problems for Co-evolution without Amplification

First, adaptation in this model is slow: adaptation in the sample learner model takes roughly an order of magnitude longer than in the No Learning No Culture Model, and
roughly two orders of magnitude longer than in the MAP model. Second, in this model, linguistic universals and strong innate constraints generally go together: with the exception of the scenario involving massive data and nativised initial conditions, linguistic universals depend upon the presence of strong innate constraints. Strong innate constraints rule out plasticity: sample learners with strong innate constraints could not easily acquire their disfavoured language. This evolutionary model provides no solution to the problem of plasticity discussed in chapter one. Third, to whatever extent stronger constraints on learning are plausibly associated with more genetic machinery, cognitive adaptation in the sample learner model predicts many associated genetic changes, which is at odds with what is known about genetic adaptations to cultural traits (Laland et al., 2010). Finally, evolution in this model fails to provide any resolution to Christiansen and Chater’s (2008) problem of dispersion: where innate constraints on cognition evolve in this model, they become strong, and thus would give rise to visible differences between long-term isolated groups.

3.3.1.3 Relation to the Moving Target Argument

That cultural transmission compromises cognitive adaptation is a familiar claim. In chapter one I discussed the proposals put forward by Christiansen and Chater (2008) in depth. They claim that cultural behaviours which vary independently of learners’ cognitive biases represent moving targets that prohibit cognitive adaptation. However, they allow the viability of the adaptationist rationale for nativism about cultural behaviours that are determined entirely by cognition (Chater et al., 2009). The models I have analysed suggest that this is too lenient. The finding that cultural transmission without amplification can stifle adaptation strengthens Christiansen and Chater’s position: it questions the evolutionary rationale for nativism as traditionally conceived, and does not rely on arguments concerning the relative speeds of biological and cultural evolution.

Overall, given the dynamics of the sample learner model, evolutionary reasoning does not inherently constitute a good rationale for linguistic nativism. That approach is only sound if either: i) relevant innate constraints are assumed to be pre-existing and exapted for language, in which case the co-evolutionary process simply bolsters these biases at best, maintains or perhaps degrades their content more likely, and at worst is
largely irrelevant, or; ii) liberal assumptions are made about mutational and selective forces such that the sub-PG problem can be overcome. The adaptationist evolutionary rationale for linguistic nativism, which equates the evolution of specialised language cognition with the evolution of specialised cognition for a-cultural behaviours, is on shaky ground according to this model. Even granting nativist and adaptationist assumptions, excluding the amplification effect of culture, and making no assumptions about the relative speeds of biological and cultural evolution, the inclusion of cultural transmission gives rise to co-evolutionary dynamics that fundamentally compromise the adaptationist rationale for the strong nativist thesis I described in chapter one.

3.3.2 Evolution, Nativism, & Culture with Amplification

Cultural transmission among populations of MAP learners results in amplification of their cognitive biases. This phenomenon radically alters the outcome of cognitive evolution. It gives rise to an evolutionary dynamic that, rather than prevent cognitive adaptation, makes that process strikingly faster and more robust. At the same time, this dynamic defuses the evolutionary rationale for strong linguistic nativism as traditionally conceived, since it never results in the fixation of strong domain-specific innate cognitive biases. Rather, it provides strong support for cognitive theories that appeal to domain-specific but weakly constraining inductive biases; evolutionary reasoning supports the *minimal nativism* discussed in chapter one. I will explain these dynamics by appeal to the concepts of evolutionary *masking* and *unmasking* (see e.g. Deacon, 2007, 2010; Wiles et al., 2005).

3.3.2.1 Masking & Unmasking

Masking and unmasking are terms that describe complementary processes which can occur in complex evolutionary systems: though they have yet to be widely discussed, these processes have: received some formal analysis (e.g. Wiles et al., 2005); been argued to be crucial in the evolution of language-related cognition (e.g. Deacon, 2007, 2010); and featured in recent models of language (Kirby et al., 2007; Smith and Kirby, 2008) and birdsong evolution (e.g. Ritchie and Kirby, 2007). Furthermore, there is empirical evidence to implicate these processes in the degradation of cognitive biases that underpin learned birdsong (e.g. Takahasi and Okanoya, 2010). It is useful to begin with a general formulation of the concepts. Variation in some variable X being *masked* by
some process P means that, where in the absence of P variation in X would lead to reproductive asymmetries, the presence of P ensures that variation in X leads to lesser reproductive asymmetries or reproductive neutrality. Unmasking denotes the opposite effect: some process P ensures that variation in X leads to reproductive asymmetries that would be lesser or non-existent in the absence of P.

In the models described above, natural selection adjusts the frequencies of genotypes. However, the reproductive fate of genotypes is determined by the phenotypes they partly cause: natural selection acts over phenotypes (languages) not genotypes. Where the effects of cognitive biases are amplified in culture, populations of learners with weak biases nevertheless converge on the favoured language. In other words, whether a learner has a weak or a strong inductive bias, he is overwhelmingly likely to end up inducing the favoured language. This means that strength of bias is masked from selection: if selection only cares what language the learner eventually induces, then it cannot see any difference between learners who possess weak or strongly constraining cognitive biases. Selection over genotypic variation is relaxed (see e.g. Deacon, 2010).

The cultural process ensures that the learner is exposed to data that largely portray the favoured language; to some extent, control of the phenotype is offloaded from genes onto culture. Because language is learned from other individuals, and the amplification dynamic ensures that those individuals tend to speak the favoured language, culture represents an organism-external informational regularity on which biological evolution can rely. All biological evolution must do to satisfy the demands of selection is mould organisms that will reliably acquire the common language given the external regularity. The masking of bias strength by amplification in cultural evolution means that, all else being equal, we have no reason to suspect biological evolution would favour any particular strength of bias over another. Without amplification, significant differences in bias strength are visible to evolution. With amplification, significant differences in bias strength are masked. Strong biases face stronger degradational pressures than weaker biases, but do not gain any fitness advantage over weaker biases. For this reason evolution in the MAP learner model never leads to strong innate constraints on learning.

Another way to interpret the effects of amplification is to consider their consequences for weak biases, or small deviations from neutrality. Without amplification, phenotypic differences between the $\alpha = 0.5$ neutral learner and the $\alpha = 0.51$ weakly biased learner are proportionally feeble, and may even be negated by mutation pres-
Chapter 3 The Evolution of Inductive Biases for Culture: A Formal Analysis

Sure: they are essentially invisible to selection. With amplification, phenotypic differences between these learners are huge, and therefore visible to selection. Absolute non-neutrality is unmasked.

Consider the dynamics of an initially unbiased population. As well as being least vulnerable to degradation through mutation, weak biases are the most likely to emerge through mutation in a neutral population; in this scenario weak biases are the most likely to emerge, the least vulnerable to degradation, and accord their possessors the maximum fitness advantage. Several consequences follow from this dynamic.

3.3.2.2 Implications of Amplification, Masking, & Unmasking

First, evolution of weak biases is extremely robust to variation in model parameters. I have found no scenarios that lead to nativisation, and only a few cases in which very accurate learning data prevents the evolution of any bias at all. Where cognitive evolution without amplification in the sample learner model was highly contingent on parameter values and initial conditions, cognitive evolution in the presence of amplification seems qualitatively impervious to these factors.

Second, adaptation, in the form of communicative coordination through a linguistic universal driven by shared cognitive biases, is rapid. Adaptation generally occurs two orders of magnitude faster than in the sample learner populations (where adaptation occurs there), and one order of magnitude faster than in the a-cultural model. Together these properties suggest that a cultural process that includes amplification actually increases the likelihood of cognitive adaptation. The model predicts an increase in the rate and number of cognitive adaptations with the onset of culture in human evolution (see e.g. Hawks et al., 2007; Laland et al., 2010, for evidence that this has occurred). At the same time, the dynamic suggests that those adaptive predispositions will be constrained to take a specific form: they will be weak. In fact, the prediction depends upon it. Cognitive adaptation in the presence of culture and amplification is fast and robust because those processes make extremely valuable a class of adaptation that is also highly evolvable (relatively easily accessible through mutation).

Third, the cognitive biases predicted by this model are defeasible: though in a cultural environment converged on the favoured variant they guarantee its acquisition, they do not compromise acquisition of alternative variants in the way that strong biases would. Exposure to data that portray alternative variants would be sufficient to overturn the
bias: the cultural environment can easily overrule the evolved preference. This provides an explanation for the problem of plasticity: it resolves any apparent paradox in the idea that linguistic features can be universal or near universal thanks to the effects of cognitive biases, yet any individual learner could easily acquire alternative variants. Such plasticity is often treated as an adaptation for flexible learning (e.g. Baronchelli et al., 2012). The model suggests an alternative explanation: flexibility arises from a combination of masking and degradational mutation pressure.

Fourth, the model also predicts that, although they may be present and even numerous, evolved cognitive biases may be extremely difficult to detect in individuals. Several consequences follow from this. For example, any genetic or cognitive variation between populations that might result from prolonged isolation and continued evolution would be essentially invisible in individuals: weak biases are, on the terms of this model, associated with small genetic differences and are defeasible in learning. Though their effects may be visible in population-level linguistic distributions, genetic testing and behavioural experiments would nevertheless give the impression of uniformity. Though this may seem counter-intuitive, it has been argued to be true of many traits: phenotypic variation often has its roots in a combination of individually undetectable small-effect genes (Rockman, 2012). The model therefore offers an explanation for Christiansen and Chater’s (2008) problem of dispersion: even if long-term isolated populations had evolved local biases, we shouldn’t expect them to be easily detectable. The biases predicted by this model might also be particularly difficult to detect in behavioural experiments. Dispositions so weak might require sophisticated bias recovery analyses (e.g. Culbertson and Smolensky, 2012; Reali and Griffiths, 2009), and or laboratory experiments that reveal inductive biases by instantiating cultural transmission experimentally (see e.g. Griffiths et al., 2008, for a rationale) or employ techniques of analysis that make predictions about long-term transmission on the basis of an experimental learning task. This may seem to be a particularly worrying suggestion: domain-specific cognitive biases may be numerous, but our experiments are powerless to find them! I will take up this issue, and argue that consequences for experimental research implied by the predictions of this model are promising rather than worrying, in chapter five.

In its solution to the problem of dispersion, does the model predict cognitive differences between isolated groups? Not in any meaningful way. Put simply: the kinds
of biases predicted by this evolutionary model are meaningless at the level of the individual, only having any observable effects when shared among populations of individuals who engage in cultural transmission for several generations; these biases can quickly shape bodies of shared knowledge over time (Dediu, 2008; Griffiths and Kalish, 2007; Kirby et al., 2007; Smith and Kirby, 2008; Dediu, 2011; Boyd and Richerson, 1985), but do not constrain individual learning in the face of counter-evidence. Though it is widely thought that all populations share the same genetic underpinnings for language, there is some evidence for between group genetic variation related to language: variation whose functional consequences are as yet unknown (Barbujani et al., 1994; Belle and Barbujani, 2007, e.g.), and variation that correlates with the phonemic use of tone (Dediu and Ladd, 2007). While there are good reasons to be cautious about large scale correlation analyses (Roberts and Winters, 2013), if these observations turn out to be genuine, the model gives a parsimonious explanation for why they have only emerged through large-scale population analyses.

3.3.3 Relation to Existing Claims

Though the models I have presented are, like all efforts in this field, abstract and highly idealised, there is growing acknowledgement (e.g. Claidière et al., 2014) and empirical evidence (e.g. Laland et al., 2010; Aoki and Feldman, 1991; Gialluisi et al., 2013) that their qualitative insights can have great explanatory potential. Whilst it is generally uncontroversial that humans do possess cognitive biases that appear specialised for learning from cultural systems (see e.g. Mesoudi, 2009 for a review of laboratory studies, and e.g. Henrich and Broesch, 2011 for field studies), is there evidence of co-evolutionary processes leading to the kinds of bias predicted by this model? Empirical estimation of biases that have emerged through gene-culture co-evolution are rare. However, there is at least one relevant case, which concerns the genetic variation associated with biases in human handedness (Laland et al., 1995). That study concludes that human handedness is a facultative trait: variation in handedness is not associated with genetic variation, but occurs because the genetic component favouring right handedness is shared among human populations but only encodes a weak bias in that direction.

The basic result that an evolutionary perspective predicts weakly biasing architectures is in line with previous suggestions. Previous results in the Bayesian iterated learning paradigm are the most obvious precedents (Kirby et al., 2007; Smith and Kirby,
2008). My arguments should be seen as extensions of ideas developed in those papers, and accordingly support similar conclusions. However, there is at least one example of a similar idea being formulated from a quite different perspective. The basic result bears out a hypothesis put forward informally by Clarke (1993), in his book *Associative Engines*. Though Clarke is concerned principally with neural networks and developmental processes, he predicts that evolution will favour weak biases based on strikingly similar logic: 

"...rather quite small initial biases, of a wide variety of kinds...can have profound implications for the way a network develops in a given environment. This "magnifying glass" effect provides an opportunity which evolution, being the laziest of designers... could not be expected to ignore. Evolution will surely favour a minimal nativism... evolution will, in the laziest way possible, have slightly loaded the dice." (Clarke, 1993, pp.185)

The "magnifying glass" effect refers broadly to the idea that environmental and developmental factors can achieve the same amplification effect as cultural transmission can achieve. That Clarke reaches the same conclusion from a different perspective is testament to the fact that the idea is not dependent upon strong theory-internal assumptions. Where some outside factor ensures that weak biases can lead to universally acquired behaviour, and some principle of cost applies to cognitive biasing, the weakest biases are optimal.

In his book *Necessary Knowledge*, Plotkin (2007) argues that minimal nativism about any learned behaviour is a necessary assumption. Plotkin argues for the general case that all learning is constrained by specialisations that are necessarily at least weakly biasing. His perspective is similar to that I have argued for here:

"...necessary knowledge ... [refers] ... to some probably very small, perhaps fragmentary, innate representational knowledge. Such necessary knowledge is the embodiment of the relationship between evolution on the one hand, and the surrogate knowledge-gaining process, i.e. learning, to which it has given rise, on the other. What this minimal innate representational knowledge does is constrain, and hence aim or guide, learning towards specific features of the world that must be learned. Such necessary knowledge

---

22In addition to Clarke's hypothesis, I am aware of one other, even more informal formulation: Shimon Ulman is reported to have floated the idea over dinner at the 35th annual conference of the cognitive science society in Berlin!
converts a potentially slow and wholly inefficient process to one that is appropriately ‘fast and frugal’ in Gigerenzer’s words” (Plotkin, 2007, pp. 249).

However, Plotkin’s arguments for necessity do not appeal to any specific evolutionary model: “The ‘necessary knowledge’ of this book owes nothing to the precise form and causes of the transformation of species in time” (Plotkin, 2007, pp. 14). One way to interpret the findings I have presented here is that, given the amplifying effect of culture, co-evolutionary modelling of specific processes shows that minimal nativism can also be sufficient.

In summary, evolutionary reasoning about a cultural trait that is subject to amplification does not support traditional nativism about that trait: in no circumstances do the models I have explored converge on shared strong biases or hard constraints. However, the analyses have revealed dynamics that also question the credibility of pure empiricism about these traits: with amplification, domain-specific cognitive adaptations in the form of weak inductive biases are a robust, rapid, outcome. This conclusion seems problematic: how can nativism and empiricism both be wrong? This reflects problems with the traditional definitions for these perspectives. In the next chapter, I will show how these results suggest a reconciliatory way to understand nativism and empiricism that preserves the most important commitments of both. The evolutionary pattern through which weak biases arise has considerable explanatory potential. Co-evolutionary dynamics under these assumptions are subtle and rich, and warrant equally subtle assumptions about the cognitive outcomes of the process.

Overall, the models I have constructed and analysed in this chapter encourage caution in accepting the evolutionary rationale for strong nativism about culturally transmitted behaviours: in a c-induction problem, where knowledge is transmitted culturally through expression, inference, and induction, and graded cognitive types are available to evolution but face proportional costs, co-evolution rarely leads to nativisation, but can in a wide range of circumstances lead to the fixation of shared defeasible cognitive biases. Of course, many of the assumptions captured by these models are open to criticism, and their relaxation may lead to different conclusions. In the next chapter, I pursue this possibility by analysing a range of variations on the models presented here.
Chapter 4

Questioning the Basics: Alternative Co-evolutionary Analyses

In the previous chapter I presented and analysed a co-evolutionary model designed to explore how culture can shape the evolution of cognition. That model was built upon a range of assumptions and idealisations whose credibility might reasonably be questioned. Here I explore a range of variations on and additions to the model and ask: do the conclusions drawn in the previous chapter still hold true under alternative assumptions? These analyses are organised into three sections that represent the core components of the co-evolutionary system: biological evolution (section 4.1), individual learning (section 4.2), and cultural transmission (section 4.3). In each analysis, the basic structure of the main co-evolutionary model should be assumed: all except the elements in focus remain as specified in the previous chapter\(^1\). Since exhaustive exploration of the parameter space for all models is impractical, I will concentrate on the default case (stronger costs to specialisation, impoverished linguistic observations), and explicitly note any deviations from this parameter regime.

\(^1\)That is, modifications to the model are not cumulative: after each analysis, any changes to the model are flushed, and the original structure should be assumed.
4.1 Biological Evolution

4.1.1 Non-Arbitrary Languages (Asymmetric Fitness)

The models presented so far include the assumption that both language types afford their users equal fitness benefits: neither language serves the communicative function better than the other. Arbitrary linguistic universals are central to nativist reasoning: many of the universals identified by linguists have no clear functional motivation (see e.g. Christiansen and Chater, 2008, for discussion of this point). However, a prominent alternative perspective assumes that particular linguistic variants offer functional advantages over others, and that universals reflect these functional considerations (see e.g. Bybee, 2009): certain linguistic structures are universal because they better satisfy communicative goals, by being well suited to the requirements of pragmatics, discourse, or processing, for example. That functionally superior linguistic variants can become universal is often explained by appeal to cultural evolution (e.g. Kirby, 1999). Crucially, however, it has also been suggested that adaptationist accounts are particularly well suited to explain linguistic universals that reflect functional considerations: "...innate autonomous grammatical principles were selected for because they allotted a greater evolutionary advantage to populations that had them." (Newmeyer, 1991, pp. 6). Existing co-evolutionary models of language appear to support this hypothesis (e.g. Kirby and Hurford, 1997; Briscoe, 2000). The distinction between functionally arbitrary and non-arbitrary aspects of language has become a significant line in the sand for critiques of adaptationist explanations for language-related cognition. For example, despite critiquing the evolutionary plausibility of nativism, Christiansen et al. (2011) explicitly endorse the adaptationist rationale for nativist accounts of functionally non-arbitrary aspects of language: they allow that evolution can build strong innate constraints so long as those constraints favour functionally superior variants.

In this section, I test this hypothesis in the model I have derived. I explore the dynamics of a model in which grammar $L_1$ is assumed to be the communicatively superior variant, and therefore affords its users a fitness advantage over users of $L_0$. For example, variant $L_1$ might be more parsable in communicative interactions, or minimise cognitive load (see e.g. Kirby and Hurford, 1997; Briscoe, 2000, for similar models).

2See for example Newmeyer (1991) or Kirby (1999) for comprehensive discussions of this perspective and how it relates to cultural and biological evolutionary considerations.
Allow the following small but significant change to the main co-evolutionary model. Assume that, for a learner with a given genome, the fitness benefit gained in the event that the learner acquires language $L_1$ (previously directly proportional to the frequency of $L_1$ among the learner’s conspecifics, $x_1$) is now multiplied by coefficient $\pi$. Where $\pi = 1$, the model reduces to the main co-evolutionary model outline above. Where $\pi > 1$, speakers of $L_1$ gain an advantage over speakers of $L_0$. For example, if $\pi = 10$, speakers of $L_1$ gain a ten-fold fitness advantage over speakers of $L_0$. In the new model, fitness is now as follows:

$$f_i(x^-, x) = \sum_{y=0}^{N} P(y|x^-) \left( [P_{learn}(L_0|y) x_0] + [P_{learn}(L_1|y) x_1 \pi] \right). \quad (4.1)$$

### 4.1.1.1 MAP Learners & Fitness Advantages

Functional asymmetries make no difference to the outcome of evolution in populations of MAP learners, except to say that they ensure that the functionally superior variant wins out over the functionally inferior variant, even if the functionally inferior variant is initially more common or nativised. In MAP populations, learners with (virtually) any bias in the direction of the common language will reliably acquire that language. Since the fitness asymmetries implied by functional asymmetries depend on the language eventually induced, and not directly on the cognitive bias, those advantages are common to weakly and strongly biased learners alike. Since weaker biases are subject to weaker mutation costs, they still out-compete stronger biases and are fixed in the population - whatever the magnitude of the functional asymmetry. Figure 4.1 shows evolution in MAP learner populations in which the functionally inferior variant is initially more common (and in the lower panel nativised from the start), and learners of $L_1$ gain a (ludicrous) one-thousand fold fitness advantage over learners who share $L_0$. Evolutionary outcomes are qualitatively identical to those in the model with functionally arbitrary languages.

---

3I am using $\pi$ as a variable, rather than the number.
Figure 4.1: Evolution in MAP learner populations who receive a 1000-fold fitness advantage ($\pi = 1000$) for sharing language $L_1$. **Left hand column panels:** distributions of genome classes. As before, the top row corresponds to Binomial initial conditions, the middle Neutral, and the bottom Nativised. **Right hand column panels:** the green line shows the mean prior bias ($\Phi$) in the population; one standard deviation in the distribution of biases in the population is shown above and below the green line by the shaded area; the purple line shows $x_1$, the population frequency of language. $n = 100$, $\mu = 0.001$, $N = 2$, $p = 0.95$
Even where the inferior variant is initially nativised (the bottom right panel of figure 4.1), weak biases for the superior variant are eventually fixed in the population: mutation pressure degrades the initial innate hard constraint towards neutrality; at that point the occasional mutant emerges with a weak bias in the direction of the superior variant; those mutants gains huge fitness advantages and take over the population, driving the superior variant to fixation.

Amplification masks bias strength from selection: it breaks the link between bias magnitude and reliable acquisition of the favoured language. Therefore, it also defuses any advantage otherwise enjoyed by stronger biases and hard constraints that favour a functionally superior language. The model suggests that, where culture leads to the amplification effect, functional considerations do not justify nativism, contrary to the claims of Christiansen et al. (2011), for example.

4.1.1.2 Sample Learners & Fitness Advantages

Figure 4.2 shows evolution in sample learner populations where learners of language $L_1$ gain a two-fold fitness advantage ($\pi = 2$). Consistent with previous models of similar processes (e.g. Kirby and Hurford, 1997; Briscoe, 2000), evolution fixes among the population strong cognitive biases favouring the functionally superior variant $L_1$. In all cases $\Phi^* = 0.88$ and $x_{1}^* = 0.9$. All models converge in less than 10000 generations.

In fact, the same qualitative outcome occurs given only a modest fitness advantage to $L_1$. Figure 4.3 shows the stable mean cognitive bias $\Phi^*$ and corresponding proportion of language $L_1$ users $x_{1}^*$ for a range of values of $\pi$. Where $\pi > 1.15$, strong innate constraints in favour of the functionally superior language are fixed in the population.
Figure 4.2: Evolution in sample learner populations who receive a fitness advantage for acquiring language $L_1$. $n = 100$, $\mu = 0.001$, $N = 2$, $p = 0.95$
Figure 4.3: Mean bias $\Phi^*$ and proportion of language $L_1$ users $x_1^*$ in the equilibrium distributions for populations of Sample learners who receive a fitness advantage for acquiring language $L_1$ proportional to $\pi$, over a range of values for $\pi$. $n = 100$, $\mu = 0.001$, $N = 2$, $p = 0.95$, Binomial initial conditions.
In the previous chapter I showed that evolution of strong innate constraints in the sample learner model can happen relatively quickly once a threshold has been crossed: in conditions that lead to nativisation, sample learner populations slowly and gradually overcome the sub-PG problem to evolve weak biases, but once a threshold bias in the population has been reached, evolution quickly builds on those biases. A similar pattern is visible in figure 4.2: gradual deviation from neutral suddenly gives way to evolution of strong constraints (from Binomial and Neutral initial conditions). Functional asymmetry seems to overcome the sub-PG problem, but only after long periods of painfully minor improvements. Stochastic simulations of the previous model showed that dynamics of this kind may be possible in the idealised population model, but not in finite populations. This possibility is worth testing here too.

4.1.1.3 Finite Populations & Fitness Advantages

Figure 4.4 shows the results of 100 simulations (50 sample, 50 MAP learner simulations) of co-evolution among finite populations (\(N = 100\)) who receive a ten-fold (\(\pi = 10\)) fitness advantage for sharing language \(L_1\). Results for MAP learner populations (purple markers) replicate the numerical analysis of the model: evolution of weak biases toward the functionally superior variant, which becomes universal. However, the outcome of evolution in sample learner populations is markedly different to that observed in the numerical analysis of idealised populations: in finite population, even a strong fitness advantage is not sufficient to scaffold the evolution of strong innate constraints in sample learners. Though the majority of simulations do converge on populations whose bias favours the functionally superior variant (\(\Phi^* > 0.5\)), those biases are weak and do not support a linguistic universal. Similar outcomes obtain for even larger fitness advantages.
4.1.2 Fitness from Learning not Communication

The fitness function of the main model rewards communicative parity at generation \( t \), an overwhelmingly common assumption in models of this kind (e.g. Nowak et al., 2001; Christiansen et al., 2011; Smith and Kirby, 2008). There are of course infinitely many ways that communication could have implications for fitness. However, there are some simple and intuitive alternative functions that are also intuitively plausible.

For instance, what would be the dynamics of the model if, rather than reward coordination among fellow learners, natural selection simply rewarded successful induction of the teacher’s grammar?\(^4\) In some sense this is arguably even closer to the spirit

\(^4\)Or, more accurately, a grammar that generates the same data as the teacher’s grammar. Selection sees behaviour, not mental representations, of course. Throughout these models, I have assumed fitness rewards parity among grammars as a close approximation to rewarding the behaviour that results from grammars. This approach is common (e.g. Smith and Kirby, 2008) in these models, and simplifies the analysis by abstracting over noisy behaviour.
of nativist reasoning about the poverty of the stimulus: impoverished data compromise the child’s ability to reconstruct the generating mental representation, therefore inductive constraints that guide the child toward that representation seem evolutionarily valuable. This assumption can be worked into the mathematical model.

This requires re-defining how fitness is calculated: now fitness must be based on the probability that a learner with a given genome will learn from a teacher who speaks a given language, and the probability that the learner will go on to induce that language given its prior preferences and her observations from that teacher. Three domains of uncertainty must be accounted for: i) the probability of learning from a teacher who speaks a particular language; ii) the possible datasets the teacher could produce (and their likelihoods); and iii) the probability of inducing the correct generating language based on those datasets. Quantity i) is determined straightforwardly by the frequency of each language at the previous generation \( (x^-) \), and quantity ii) is given by the binomial likelihoods for the possible counts of rule \( S \to X Y \) given each language. Quantity iii) will differ between sample and MAP learners, since it depends upon induction probabilities. In broad terms though, I must sum over all possible dataset counts, asking: what is the likelihood that a learner would see this dataset from a teacher who speaks language \( L_l \), and if it did, what is probability that the learner would induce language \( L_l \)? Therefore fitness will be given by:

\[
 f_i(x^-) = \sum_{y=0}^{N} (\frac{N!}{y!(N-y)!}) \frac{p^y(1-p)^{N-y}}{1 + O_{L_1}^{+\infty}} x_0^- + \sum_{y=0}^{N} P(y|L_1) P_{\text{learn}}(L_1|y, |q| = i),
\]

where \( P_{\text{learn}}(L_i|y) \) stands in for the sample and MAP learner induction probabilities. In the subsections below I define fitness calculations precisely in each of these models, and analyse the resulting co-evolutionary dynamic.

### 4.1.2.1 MAP Learners

Induction probabilities amplify posterior probabilities in MAP learners, so that fitness in this model is given by:

\[
 f_i(x^-) = \sum_{y=0}^{N} \frac{N!}{y!(N-y)!} \left( \frac{p^y(1-p)^{N-y}}{1 + O_{L_1}^{+\infty}} x_0^- + \frac{p^{N-y}(1-p)^y}{1 + O_{L_0}^{+\infty}} x_1^- \right),
\]

where

\[
 P_{\text{learn}}(L_i|y) \text{ stands in for the sample and MAP learner induction probabilities.}
\]
where, as before, posterior odds are given by:

\[
O_{L_0} = \frac{n - i}{i} \left( \frac{p}{1 - p} \right)^{2y-N}
\]

(4.4)

\[
O_{L_1} = \frac{i}{n - i} \left( \frac{1 - p}{p} \right)^{2y-N}
\]

(4.5)

Evolutionary outcomes are in this model identical to those observed for the basic model, for all conditions I have tested. For instance, in the default scenario \((N = 2, p = 0.95, n = 100, \mu = 0.001)\) where MAP learners receive fitness directly proportional to the probability that they will acquire the language from which their utterances were generated, evolution rapidly fixes weak biases in the population \((\Phi^* = 0.55)\), and these biases give rise to a linguistic universal \((x_1^* = 0.94)\), from all three sets of initial conditions. Since the initial/final distribution and time-course plots are identical to those from the basic model, I omit them here.

### 4.1.2.2 Sample Learners

Since for the sample learner induction probabilities follow posterior probabilities, fitness is calculated, after some simplification, with:

\[
f_i(x^-) = \sum_{y=0}^{N} \binom{N}{y} x_0^{2y} p^{2y}(1-p)^{2N-2y} (n-i) + x_i^-(1-p)^{2y} p^{2N-2y} i.
\]

(4.6)

Unlike the MAP learner model, the sample learner model in which fitness gains are accrued through accurate learning demonstrates subtly different behaviour to the basic version in which fitness is based on coordination with the current generation. For instance, figure 4.5 shows evolution in this model under the assumptions most favourable to nativist reasoning (lower costs to specialisation and lower quality data: \(n = 50, N = 2\)). Recall that the these assumptions facilitated the the evolution of strong innate constraints from all initial conditions in the basic model. However, in this model, all conditions lead to neutrality: even where one language is initially nativised, those strong constraints are degraded through mutation until the population settles on the distribution defined by mutation pressures.
I have found no conditions that lead to nativisation in the sample learner model with this fitness regime. In this scenario, bias is only useful when learning from a teacher who speaks the favoured language. Without amplification, sample learners are likely to encounter teachers who speak either of the languages. Even when one language is universal, bias doesn’t bring with it enough fitness gain to overcome mutation pressure: learning about the teacher is an easier task than coordinating with the current generation, since utterances provide direct evidence about the target grammar. For these reasons, bias doesn’t seem to be evolvable or maintainable in sampler learner populations under this fitness regime.
4.1.3 Continuous Biological Types

In the main co-evolutionary model, the set of possible genomes was finite. Here I relax that assumption, discarding the bit-string genome model in favour of allowing biological types to vary continuously. Since the numerical analysis of the mathematical model requires a finite set of biological variants to track, that approach is not simple to implement here. As such, I’ll re-define the genome model and simulate the resulting co-evolutionary system in finite populations. Simulations operate in a population \( A = (a_1, a_2, \ldots, a_N) \) of \( N \) agents, where each agent \( a_i = (g_i, l_i) \) is represented by a genome \( g_i \in [0, 1] \) and a language \( l_i \in \{L_0, L_1\} \). Now an agent’s genome \( g_i \) is simply a real number on the unit interval, and specifies the learning bias directly, so that \( \alpha_i = g_i \).

In order to be able to capture assumptions about the relative probability of genetic types, I require a measure over the domain of the prior: the interval \([0, 1]\). The family of Beta distributions provides a continuous measure over this space. In this model, I treat \( g \) as the heritable biological unit, and assume \( g \sim \text{Beta}(\beta, \beta) \). Where \( \beta > 1 \), we obtain a symmetric bell-shaped distribution centred at 0.5. This encodes the assumption that the stronger biases are biologically least probable. As \( \beta \to +\infty \), the distribution becomes more peaked. The bit-string genome model enforced the Binomial distribution over the integers \( 1, 2, \ldots, n \), which were mapped into the \([0, 1]\) space with the function \( i \mapsto i/n \), \( i \in \mathbb{Z}, n \in \mathbb{Z}_1 \). I concentrated on the case \( n = 100 \). A very close approximation to this distribution is obtained with the Beta distribution by setting \( \beta = 50 \). This allows a direct comparison of the continuous and discrete cases.

The workings of mutation must also be re-specified. A mutation operator is required that: 1) favours small jumps between parent and mutant; 2) allows a continuous distribution over mutants, and; 3) implements the Beta distribution just described. Computationally, mutation is implemented with a rejection sampling scheme. Where a mutation occurs, a candidate mutant is drawn from a normal distribution with mean \( g \) and standard deviation \( \sigma \). This satisfies requirements 1) and 2). If the mutant lies outside the legal range \([0, 1]\) it is reflected back in: for example, if a candidate mutant lies some distance \( \epsilon \) above the upper boundary 1, the reflected value lies an equal distance inside that boundary at \( 1 - \epsilon \) (Bullock, 2001). Though this step seems counter-intuitive, it has been shown (Bullock, 2001) to be necessary to ensure the mutation operator doesn’t bias the distribution of mutants towards the extremities of the legal range. The candidate mutant \( g' \) is then accepted with probability proportional to \( \text{Beta}(g'; \beta, \beta) \), or re-
jected with probability proportional to $1 - \text{Beta}(g'; \beta, \beta)$, satisfying requirement 3). The process repeats until a candidate is accepted.

Figure 4.6 shows the outcomes of 100 simulations (50 MAP, purple markers & 50 sample, yellow markers) of evolution in this model under conditions directly comparable ($\beta = 50, \mu = 0.001, \sigma = 0.01$) to the default set-up in the original genome model (where $n = 100$). Shown in that figure by each marker is the mean cognitive bias in the population, $\Phi = \frac{1}{N} \sum_{i=1}^{N} g_i$, and the proportion $x_1$ of agents speaking language $L_1$, after 5000 generations. Results appear qualitatively identical to those observed under basic model: fixation of weak biases in MAP populations, and random variation around neutral in sample learner populations.

Figure 4.6: Cognitive and linguistic outcomes after 5000 generations of co-evolution in populations of $N = 100$ sample learners (yellow markers) and MAP learners (purple markers) whose genomes are drawn from a continuous range of possibilities. $\beta = 50, \sigma = 0.01, \mu = 0.001, N = 2, p = 0.95$
This genetic model allows exploration of the consequences that result from assumptions about the relative costs to specialisation\(^5\) without restricting the range of possible genetic types, by varying the cost parameter $\beta$. Figure 4.7 shows the distribution over genetic types defined by the Beta distribution for a range values for $\beta$.

![Figure 4.7: Example beta distributions.](image)

For instance, figure 4.8 shows evolutionary outcomes where specialisation is not costly: all biases are equally probable ($\beta = 1$).

In this case evolutionary reasoning does support nativism about populations of sample learners, who generally converge on shared strong biases that drive a linguistic universal, with some variation in strength. Cognitive outcomes in MAP learner populations are distributed randomly among the biases matching the linguistic universal: for the MAP learner, any bias whose direction is shared among the population is good enough to achieve maximum fitness. Even without the assumption of cost to specialisation, evolution does not favour stronger biases over weaker biases in MAP populations: the strongest viable evolutionary prediction is that selection for communication favours shared non-neutrality over neutrality.

\(^{5}\)This is an analogue of the assumption about cost in the bitstring model: the cost imposed by asymmetries in the prior probabilities of different genomes (more strongly biasing genomes are less probable a priori), and implemented through mutation.
Figure 4.8: Cognitive and linguistic outcomes after 5000 generations of co-evolution in populations of \( N = 100 \) sample learners (yellow markers) and MAP learners (purple markers) whose genomes are drawn from a continuous range of possibilities, and face no cost to specialisation. \( \beta = 1, \sigma = 0.01, \mu = 0.001, N = 2, p = 0.95 \).

Figure 4.9 shows equivalent plots for simulations in which costs to specialisation were weak, for two of the beta distributions shown in figure 4.7. Lower costs (left, \( \beta = 2 \)) reduce the proportion of sample learner populations that reach strong innate constraints, but do not prevent that outcome, and have little effect on MAP learner populations. Stronger, but still relatively weak, costs to specialisation (right, \( \beta = 10 \)) already cap the strength of the biases evolvable in sample learner populations, and recover the MAP learner results observed in previous models.
Chapter 4  Questioning the Basics: Alternative Co-evolutionary Analyses

Figure 4.9: Evolutionary outcomes after 5000 generations under weak costs to specialisation in sample learner (yellow points) and MAP learner populations, for two of the beta distributions shown in figure 4.7. Panel (a) shows the outcome of simulations with very weak ($\beta = 2$) costs, while panel (b) shows those outcomes under stronger, but still relatively weak, costs ($\beta = 10$). $N = 100, u = 0.001, \sigma = 0.01, N = 2, p = 0.95$

4.2 Individual Learning

4.2.1 Interpolation between MAP & Sample

MAP and sample learning can be viewed as opposing extremes on a continuum of hypothesis selection strategies. If induction probabilities are proportional to posterior probabilities exponentiated by some exponent $r$, so that $P_{\text{learn}}(L_l|y) \sim P(L_l|y)^r$ (Kirby et al., 2007), then the case $r = 1$ encodes no amplification of posterior probabilities and corresponds to sample learning, and, in the limit, $r \to +\infty$ corresponds to MAP learning. It is simple to accommodate this change in the mathematical model: rather than assume separate MAP and sample learning models, here I assume all learners compute posterior probabilities and transform those into induction probabilities by exponentiating the distribution by $r$ and re-normalising, so that:

$$P_{\text{induce}}(L_l|y, |q| = i) = \frac{P(L_l|y, |q| = i)^r}{P(L_0|y, |q| = i)^r + P(L_1|y, |q| = i)^r} . \quad (4.7)$$

This assumption leads to some surprising behaviour in the model. For instance, given the default conditions (Binomial initial conditions, $n = 100, \mu = 0.001, N = 2, p = 0.95$), where (roughly) $r \geq 7$, evolutionary dynamics and outcomes are essentially equivalent to those observed in MAP learner populations: rapid fixation of weak biases that sup-
port a corresponding linguistic universal. Stronger exponentiation speeds up adaptation and leads to gradually weaker biases. Figure 4.10 shows mean bias $\Phi^*$ and corresponding proportion of $L_1$ users $x_1^*$ in the equilibrium distributions reached for a range of values for $r \geq 7$.

Figure 4.10: Mean bias $\Phi^*$ and proportion of language $L_1$ users $x_1^*$ in the equilibrium distributions for populations of learners whose induction probabilities follow posterior probabilities exponentiated by $r$, over a range of values for $r$. $n = 100$, $\mu = 0.001$, $N = 2$, $p = 0.95$, Binomial initial conditions.

However, where $r \leq 7$, the model does not converge to a single stable distribution of genomes (or languages), but rather converges to a limit cycle: an apparently sustainable oscillation between distributions. Limit cycles are known to be possible in the replicator-mutator equation (Pais and Leonard, 2011), and such chaotic behaviour has been discovered in language models based on the replicator dynamics (Mitchener and Nowak, 2004) and in gene-culture co-evolutionary models (e.g. Kendal et al., 2009) before. Figure 4.11 shows a time course plot for the case $r = 2$. 

162
Figure 4.11: Evolution in populations of learners whose induction probabilities follow posterior probabilities exponentiated by $r = 2$. The green line and fill show the mean cognitive bias $\Phi$ and half of one standard deviation in that distribution. The purple line shows the proportion of the population speaking language $L_1$. $n = 100$, $\mu = 0.001$, $N = 2$, $p = 0.95$, Binomial initial conditions.

Co-evolution gradually settles into an oscillation between reasonably strong biases ($\Phi \approx 0.85$) and reasonably weak biases ($\Phi \approx 0.7$). Though the dynamics of this system are difficult to interpret, the following account is reasonably intuitive. Posterior exponentiation causes a non-linear relation between bias strength and acquisition probabilities. When the majority language (here $L_1$) is less frequent, stronger biases enjoy a more than proportional advantage over weaker biases due to imposing an exponentially greater probability of acquiring that language given data that imply it. As the population frequency of these stronger biases grows, so does the population frequency of $L_1$. However, once $L_1$ is more common in the population, the exponential advantage to stronger biases is diluted, since learners with weaker biases are also more likely to learn from speakers of $L_1$, and are therefore more likely to induce that language despite their weaker bias. As the frequency of weaker biases increases at the expense of stronger biases, the frequency of $L_1$ decreases until stronger biases regain their advantage, and the dynamic repeats this cycle. Greater exponentiation increases population level amplification, decreases the frequency and distance of the oscillation, and decreases the average strength of the average bias. Figure 4.12 shows an equivalent plot.
for \( r = 5 \).

![Graph showing evolution in populations of learners](image)

Figure 4.12: Evolution in populations of learners whose induction probabilities follow posterior probabilities exponentiated by \( r = 5 \). The green line and fill show the mean cognitive bias \( \Phi \) and half of one standard deviation in that distribution. The purple line shows the proportion of the population speaking language \( L_1 \). \( n = 100, \mu = 0.001, N = 2, p = 0.95 \), Binomial initial conditions.

Overall, any amplification of posterior probabilities leads to some population level amplification of innate preferences. In spite of the complex dynamics, even weak posterior exponentiation seems to overcome the \textbf{sub-PG} problem and facilitate the evolution of (on average, fairly unconstraining) inductive biases, but never leads to the fixation of strong constraints.

### 4.2.2 MAP vs. Sample Learners

Given the huge disparity between results for MAP and sample learners, it is natural to ask how evolution might select among those types of learner if both were simultaneously present in the population. Throughout the thesis, I have contrasted these two types of learner as a proxy for learning mechanisms that do and do not give rise to amplification over transmission. When both types of learner are competing in the population, evolution is given a choice between those two kinds of culture: this is a crucial
analysis, since it speaks to the plausibility of the distinct cognitive outcomes associated with each co-evolutionary pattern. Smith and Kirby (2008) showed that evolution tends to favour MAP learners over samplers. This prediction is borne out in the current models too: figure 4.13 shows time-course graphs for typical simulations of the basic co-evolutionary model in finite populations that contain both MAP and sample learners, over a range of assumptions about the initial composition of those populations. Here each learner inherits a single gene that controls whether she uses the MAP or sampling hypothesis selection strategy. That gene is inherited from the biological parent along with the set of genes that determines the prior, and is subject to symmetric mutation at the same rate $(\mu)$. Even when sample learners make up the whole population initially (panels A-C in figure 4.13), MAP mutants quickly take over, and the resultant evolutionary pattern reflects the amplification dynamics associated with MAP learners. In all scenarios the proportion of sample learners in the population (the purple line) approaches 0, the population converges on weak non-neutral biases (the red line, showing $\Phi$ settles just above or below 0.5), which supports a linguistic universal (the blue line, which gives $x_1$, converges to near 0 or near 1).

In line with Smith and Kirby’s predictions, evolution in this model favours MAP learners. The model predicts that evolution will favour learning mechanisms that give rise to amplification in cultural contexts over those that don’t.

Whether the MAP or sampling model is a better characterisation of human learning is an empirical question that may well vary by domain. One way to think about the differences between these strategies is in terms of cooperation, or trust. Sample learners hedge their bets, entertaining the induction of competing hypotheses in proportion to their subjective beliefs about those hypotheses. Such a strategy has clear benefits in certain scenarios: under high uncertainty, for instance, or in problems where erroneous inferences about some learner-independent truth could be costly (problems of n-induction). Experimental studies have shown sample learning to be a good model of human cognition (e.g. Reali and Griffiths, 2009), with rational motivation (e.g. Denison et al., 2013; Bonawitz et al., 2014), in at least some domains. MAP learning, or posterior exponentiation, on the other hand, corresponds to the notion of trusting one’s instincts, and has clear value when learning from other similar learners (problems of c-induction), where parity takes precedence over accuracy: among similarly biased minds, MAP learning is a good strategy for coordination problems.
Figure 4.13: Evolution in populations that contain both MAP and sample learners. In all panels: the purple line shows the proportion of sample learners in the population; the red line shows the mean innate bias $\Phi$ in the populations, and; the blue line shows $x_1$, the population frequency of speakers of language $L_1$. Plots show differing initial conditions with respect to the initial frequency of sample learners (All-None-Half) over the rows, and the mean innate bias of the first population over the columns (Neutral-Nativised $L_1$ -Nativised $L_0$). All conditions lead to the same outcome: MAP learners with weak innate biases supporting a linguistic universal. $N = 100, n = 100, \mu = 0.001, N = 2, p = 0.95$
4.2.3 A Continuous Space of Languages

Recall the model of grammar induction laid out as part of the basic co-evolutionary model. There learners considered just two possible PCFGs: $L_0$, which favoured $S \rightarrow XY$, and $L_1$, which favoured the opposite ordering pattern. Here I extend the hypothesis space to include an infinite number of such grammars. Rather than reason about the likelihood of two pre-specified sets of rule probabilities ($\theta_0$ & $\theta_1$) as before, here I assume learners consider all possible rule probabilities. The learner must infer from her observations the probability $p$ of rule $S \rightarrow XY$. Another way to think about this model is that the learner knows her observations have been produced by a PCFG, but does not know the rule probabilities and must therefore estimate these. This model of grammar induction is widely used in natural language processing and machine learning (e.g. Johnson et al., 2007), and has been applied as a model of innate inductive biases in artificial language learning experiments (e.g. Culbertson and Smolensky, 2012). Formally, let $L$ be a PCFG of exactly the kind outlined earlier, with rule probabilities:

$$\bar{\theta} = (\theta_{S \rightarrow XY} = p, \theta_{S \rightarrow YX} = 1 - p). \quad (4.8)$$

Having observed $N$ utterances with $y$ applications of rule $S \rightarrow XY$, the learner must reason about $\bar{\theta}$ by inducing an estimate, $\hat{p}$, of the true underlying probability $p$ of rule $S \rightarrow XY$. To do so she must compute the posterior distribution $P(p|y, N)$. Again assuming Bayesian inference, this is achieved by combining the likelihood of the observations, $P(y|N, p)$, and a prior distribution $P(p)$. Likelihoods are given by the Binomial sampling model exactly as before, so that:

$$P(y|N, p) = \binom{N}{y} p^y (1 - p)^{N-y}. \quad (4.9)$$

The prior distribution $P(p)$ represents the inductive bias of the learner. Given the Binomial sampling model, an obvious choice of prior distribution is the Beta distribution. The shape of the Beta distribution dictates the prior preferences of the learner. As discussed in section 4.1.3, the uniform prior is captured by the Beta$(1, 1)$ distribution: all values of $p$ are equally likely under this distribution; a learner with this set of prior beliefs would be equally happy learning any rule probabilities. Here I model prior bias by assuming learners’ preferences are encoded by the distribution Beta$(1 + \alpha, 1)$, with the constraint that $\alpha \geq 0$. Where $\alpha = 0$, the learner is unbiased; where $\alpha > 0$, the learner
favours higher values of $p$, and is therefore biased to learn grammars that tend to consistently employ the rule $S \rightarrow X Y$. Larger values of $\alpha$ strengthen this preference, so that as $\alpha \rightarrow +\infty$, the learner is constrained more tightly to only entertain values $p \approx 1$. Small values of $\alpha$ encode a weak bias in favour of values $p > 0.5$ (see figure 4.14b for some example prior distributions over $p$ given various biases $\alpha$).

The Beta distribution is the conjugate prior for the $p$ parameter of the Binomial distribution. Conjugacy affords various mathematical conveniences, including a standard form for the posterior distribution, which in this case is:

$$ P(p|y, N) = \text{Beta}(1 + \alpha + y, 1 + N - y) $$

$$ = \frac{p^{\alpha+y}(1-p)^{N-y}}{B(1+\alpha, 1)} , $$

where $B(x, y)$ represents the standard Beta function.

As before, I explore evolution in populations of MAP and sample learners. The sample learner’s induced estimate $\hat{p}$ follows the posterior probabilities exactly, so $\hat{p} \sim P(p|y, N)$. The MAP learner will induce an estimate $\hat{p}$ that maximises posterior probability. In this model the MAP $p$ value is given by:

$$ p_{\text{MAP}} = \frac{\alpha + y}{2\alpha + N} , $$

so for MAP learners $\hat{p} = p_{\text{MAP}}$. Finally, having induced its estimate of the rule probabilities, a learner’s productions follow the Binomial likelihoods given in (4.9).

This model of language learning necessitates two further changes to the basic model. First, I must specify how the bias, $\alpha$, is inherited. The bit-string model of genomes is no longer appropriate: its properties were well suited to the domain of the prior in the original model (the unit interval), but do not map well onto the domain of the prior in this model (the positive real numbers). Second, I must specify how communicative accuracy, and therefore fitness, is calculated.

The bias parameter $\alpha$ is a positive real. Given the assumptions laid out previously,

---

6Note that this model only allows biases that favour the rule $S \rightarrow X Y$. Though this space of biases is continuous and allows graded preferences, it does not allow biases in favour of the opposite ordering pattern. This is a slightly different assumption to that of the original learning model, which allowed a symmetric set of biases that could favour either grammar.

7See e.g. Box and Tiao (1992) and Gelman et al. (2003) for overviews of the Bayesian methods I have used in this thesis.
Figure 4.14: (a) The distribution (probability density) of priors $\alpha$ under the assumption that genomes follow an exponential distribution parametrised with rate $\lambda = 5$: the distribution strongly favours the weakest biases. (b) The prior distribution (probability density) over $p$ for a range of values of $\alpha$: higher $\alpha$ implies a stronger inductive bias in favour of the ordering rule $S \rightarrow X Y$.

I require a measure over this space that 1) imposes costs on stronger biases, and can be implemented through a mutation operator that 2) ensures sensible jumps between types. As in section 4.1.3, requirement 2) can be enforced by the normal distribution with the parent $\alpha$ as its mean, and standard deviation $\sigma$. I implement requirement 1) with the exponential distribution, parametrized with rate parameter $\lambda = 5$. This distribution has mode 0 and mean $\lambda^{-1} = 0.2$. Figure 4.14a visualises the the prior probability of biases $\alpha$ implied by this distribution over genomes.

In fact, because the space of genetic types $\alpha$ is continuous, the exactly neutral prior $\alpha = 0$ is technically impossible: the distribution strongly favours the weakest biases, but ensures that all learners possess some bias. This is a slightly different assumption to that represented in the bitstring genome model, but arguably reflects an equally plausible scenario that is also worth exploring: while mutation degrades information content, no learner is every truly neutral. In brief, when a mutation occurs, the distribution over possible mutants is proportional to the product of a Gaussian around the parent and an Exponential distribution that favours the weakest possible biases. Computationally, mutation is implemented with the rejection sampling algorithm described in section 4.1.3, with the Exponential distribution replacing the Beta distribution described there.

Here as before fitness is associated with linguistic coordination: a learner’s com-
municative accuracy reflects the probability that it would, in encounters with other members of the population, employ the same linguistic variant (expansion of S) as its partners. Given a population \( A = \{a_1, a_2, \ldots, a_N\} \) of \( N \) agents, the fitness \( f_i \) of agent \( a_i \) is given by:

\[
f_i(A) = \frac{1}{N} \sum_{j=1}^{N} \hat{p}_i \hat{p}_j + (1 - \hat{p}_i)(1 - \hat{p}_j),
\]

(4.12)

where \( \hat{p}_i \) is the estimate induced by agent \( a_i \). This assumption dictates that the fittest population is one whose members all employ the same single linguistic variant. Consequently, learners who induce rule probabilities \( p \approx 1 \) will tend to be fitter, since they will consistently use \( S \rightarrow XY \). We should expect evolution to favour biases \( \alpha > 0 \), then, since this biases learners towards regular rule probabilities. The evolutionary unknown thus concerns the strength of the resultant biases. The adaptationist rationale for linguistic nativism predicts in this case that \( \alpha >> 0 \).

For each type of learner, I report the results of 100 simulations, each of which ran for 5000 generations. Figure 4.15 shows the distribution of final population mean biases \( \Phi = \frac{1}{N} \sum_{i=0}^{N} \alpha_i \) over 100 simulations of co-evolution in populations of MAP learners against the expected distribution through mutation alone\(^8\). Simulations invariably resulted in the fixation of weak inductive biases that favour higher values of \( p \). Though those biases are weak, they are marginally stronger than would be expected through mutation alone (the peak of the distribution in figure 4.15 is not on the biases nearest to zero). All simulations resulted in linguistic universals: allowing \( p^* = \frac{1}{N} \sum_{i=1}^{N} p_i \) to be the average value of \( p \) in the grammars induced by members of the final population, then \( p^* \approx 1 \). for all simulations. As such, the values are not plotted. Results in this model portray the same qualitative co-evolutionary pattern as did the analysis of the two-grammar basic model.

Figure 4.16 shows equivalent outcomes in sample learner populations. Here populations tended to settle on the weakest biases, as implied by the mutation dynamics, and did not converge on linguistic universals (the mean of \( p^* \) values across simulations is 0.52). Figure 4.17 shows the distribution of \( p^* \) values over simulations.

---

\( ^8 \)I have scaled the probability density function of the exponential distribution over genomes to be directly comparable to the histogram of simulation results: the important comparison is the shape of the distributions.
Figure 4.15: The distribution of final-population mean cognitive biases $\Phi$ over 100 simulations of 5000 generations in populations of MAP learners. Each bar gives the proportion of simulations in which the average final population bias fell within the range of the bin. The black dashed line shows the expected distribution of biases if there were no selection (scaled for easy comparison). $N = 100, \lambda = 5, \sigma = .1, \mu = 0.001, N = 4$

Figure 4.16: The distribution of final-population mean cognitive biases $\Phi$ over 100 simulations of 5000 generations in populations of Sample learners. Each bar gives the proportion of simulations in which the average final population bias fell within the range of the bin. The black dashed line shows the expected distribution of biases if there were no selection (scaled for easy comparison). $N = 100, \lambda = 5, \sigma = .1, \mu = 0.001, N = 4$
Figure 4.17: The distribution of final-population $p^*$ values over 100 simulations of 5000 generations in populations of Sample learners. Each bar shows the proportion of simulations in which the average of the final population’s $p$ values (which dictate the relative use of the two possible ordering patterns) fell within the range of the bin. $N = 100, \lambda = 5., \sigma = .1, \mu = 0.001, N = 4$

Again, for the sample learner, the evolvable weak biases are insufficient to drive strong asymmetries in the distribution of languages in the population that would create a selection pressure for stronger biases. Co-evolution never gets off the ground. Overall, these analyses illustrate an important point: the general pattern of evolutionary dynamics revealed in the main co-evolutionary model is not specific to the assumption that only two possible languages exist. Those dynamics have been recovered here, where learners must choose among a technically infinite number of hypotheses (possible probabilistic constraints on the use of the two ordering patterns). While it will no doubt be instructive for future research to explore these dynamics in models that include hypothesis spaces with more structure, this manipulation lends confidence to the generality of the results.
Chapter 4 Questioning the Basics: Alternative Co-evolutionary Analyses

4.3 Cultural Transmission

4.3.1 Multiple Teachers

In the basic model I assumed that each learner observed utterances generated by a single teacher from the previous generation. This idealisation has been widely adopted in models of cultural evolution (e.g. Kirby, 2002; Kirby et al., 2007), and gene-culture co-evolution (e.g. Smith and Kirby, 2008), and has been shown to lead to similar conclusions as population models in certain scenarios (e.g. Griffiths and Kalish, 2007). However, the assumption is commonly criticised: for instance, Niyogi and Berwick (2009) argue strongly that single-teacher cultural transmission models lack qualitative dynamics of multiple teacher models known to be important in language change, and Enquist et al. (2010) argue that single teacher transmission chains can reduce cultural variation.

Research into the dynamics of cultural transmission through iterated Bayesian learning from multiple teachers shows two clear results. First, Smith (2009) shows that cultural transmission among populations of Bayesian learners who receive data from multiple teachers leads to amplification of the prior similar to that observed in MAP learner populations. However, this model has been questioned (Burkett and Griffiths, 2010) on the grounds that its learners are irrational: each learner receives data generated by (potentially) many speakers, but reasons about the data on the basis that they were generated by a single speaker.

The second important result generalises Smith’s model to allow learners that explicitly draw inferences about the number of languages being spoken among the teaching population (Burkett and Griffiths, 2010). This estimate is made under the influence of an inductive bias that determines whether learners expect to encounter one, few, or many languages. This is important because the learner’s expectations about the degree of linguistic variation can have significant long-term consequences on the cultural distribution of languages. Where learners expect to encounter many languages, the cultural process converges on a linguistic distribution specified by learners’ prior preferences: there is no amplification. However, where learners are biased to expect less linguistic variation, transmission dynamics resemble those derived by Smith: substantive prior preferences are amplified through the process of cultural transmission. These results are crucial because they suggest another way, in addition to MAP learning, that aspects of the inferential process other than strong substantive bias can bring
about universals. Both results concern cultural transmission among cognitively homogeneous populations. Dediu (2009) has argued that the finding might not generalise to cognitively heterogeneous populations: he shows that more complex populations deliver results that resemble single-teacher sample learner models (no amplification).

The assumption that learners observe data from multiple individuals but perform inference over the dataset as if it was generated by a single speaker can be straightforwardly accommodated in the mathematical model. All that must be re-formulated are the dynamics of cultural transmission. This affects the probabilities that a learner will observe in her data a particular count \( y \) of the rule \( S \rightarrow X Y \) given that she now samples each of her \( N \) utterances from \( N \) randomly chosen (with replacement) teachers at the previous generation. In that case, the probability of observing count \( y \) given the distribution of languages \( x^- \) at the previous generation is:

\[
P(y|x^-) = \left[ (px_0^-)^y + (1 - p)x_1^- \right]^y \times \left[ (1 - p)x_0^- N-y + (px_1^-)^N-y \right]
\]

With this assumption, the model of transmission and learning is mechanically identical to the (two grammar, monolingual) case explored by Smith (2009). The following subsections report co-evolutionary dynamics in sample and MAP learner populations respectively.

### 4.3.1.1 Multiple Teachers & Sample Learners

Figure 4.18 shows evolution in populations of sample learners who learn their languages from data generated by multiple individuals. All three sets of initial conditions lead to the same outcome: fixation of moderately weak cognitive biases that support a linguistic universal. In all cases \( \Phi^* = 0.71, x_1^* = 0.93 \). The model converges to equilibrium after around 7000 generations. The linguistic universal emerges faster than this though: given Binomial and Neutral initial conditions, language \( L_1 \) goes to fixation after around 1000 generations. Amplification has again sped up adaptation (compared to the single teacher sample learner model under conditions that allowed cognitive specialisation). Note that the stable mean cognitive bias in the population is stronger than that observed in the MAP models of the previous chapter. Why is this?
Figure 4.18: Co-evolution in populations of sample learners that learn from multiple individuals. \( n = 100, \mu = 0.001, N = 2, p = 0.95 \)

Two important details about the nature of amplification in this model explain the difference. First, Smith (2009) showed that, in this model of transmission, larger transmission bottlenecks (the size of the dataset, \( N \)) are associated with an increased degree of amplification. So, for example, if a population shares the weak inductive bias (in favour of language \( L_1 \)) \( \alpha = 0.51 \), but only observes relatively impoverished data (\( N = 2 \)), language \( L_1 \) eventually settles into a stable population frequency of just \( x_1^* = 0.54 \): there is some amplification, but not much, with \( N = 2 \). Because bias amplification is weak under these conditions, the population must share a stronger bias to achieve the fitness benefits that result from a language being shared by the whole
population (i.e. $x^*_1 \approx 1$). Smith showed that $L_1$ only approaches universal status in the population (roughly, $x^*_1 > 0.9$) when the shared bias (roughly) satisfies $\alpha > 0.7$. The co-evolutionary population analysis bears out this dynamic, and fixes slightly stronger biases in the population so that $\Phi^* = 0.71$, which ensures $L_1$ goes to near-fixation, so that $x^*_1 = 0.93$.

The second important detail about the amplification dynamic in this model is that the linguistic distribution in the initial population can affect the long term distribution of languages: over transmission, the initial conditions can be amplified too. For instance, with a larger bottleneck (more amplification), if language $L_0$ were initially more frequent than $L_1$, amplification of these initial conditions could overrule a weak bias that favoured $L_1$ (but not a strong bias) and result in fixation of language $L_0$ as universal. In sum, with large bottlenecks, historical conditions can be amplified and entrenched.

Together these details of the amplification process in this model create the potential for dynamics that differ from those observed under MAP learning. Figure 4.19 shows the co-evolutionary outcomes when learners observe rich data ($N = 6$). In this case $\Phi^* = 0.503$ and $x^*_1 = .998$: the average bias fixed in the population is extremely weak, while the linguistic universal is extremely strong, and emerges extremely quickly. Though the population is ever so slightly skewed in favour of biases for $L_1$, overall it is essentially indistinguishable from neutral. With such a large bottleneck, the initial slight advantage to $L_1$ ($x^*_0 = 0.55$) has been amplified and entrenched, such that substantial cognitive biases are not required to achieve linguistic coordination.
4.3.1.2 Multiple Teachers & MAP Learners

Figure 4.20 shows evolution in populations of MAP learners who learn their languages from data generated by multiple individuals. The co-evolutionary dynamics and cognitive and linguistic outcomes here appear qualitatively identical to those observed in single-teacher MAP models. In all cases weak biases are rapidly stabilised and support a linguistic universal: $\Phi^* = 0.56, x_1^* = 0.99$. 

Figure 4.19: Co-evolution in populations of sample learners that learn from multiple individuals. $n = 100, \mu = 0.001, N = 6, p = 0.95$
Figure 4.20: Co-evolution in populations of MAP learners who learn from multiple individuals. $n = 100, \mu = 0.001, N = 2, p = 0.95$

In single-teacher models, smaller datasets (smaller $N$) increase the degree of amplification in MAP learner populations (Griffiths and Kalish, 2007). In multiple-teacher populations, however, the trend is the opposite (Smith, 2009): more data leads to more amplification, as in the sample learner model discussed above. Figure 4.20 shows co-evolution in the MAP multiple teacher model where learners observe more data ($N = 6$).
As in the sample learner model above, the larger bottleneck entrenches the linguistic universal: consequently, the pressure to be biased in favour of that majority language is weaker, and the biases that evolve are accordingly weaker: $\Phi^* = 0.52, x_1^* = 0.99$ for all starting conditions.

For both MAP and sample learners, co-evolution among populations that learn from multiple individuals results in the fixation of weak cognitive biases and the emergence of a linguistic universal. Where the data learners observe are assumed to be rich, slight initial linguistic asymmetries can be amplified into universals, leading to a subtly different co-evolutionary process that nevertheless also leads to weak biases or neutrality.
By giving rise to amplification, and therefore the co-evolutionary short-cuts to adaptation that affords, performing inference over data as if they reflect a single generating language is, evolutionarily speaking, an extremely valuable strategy. Despite being irrational, that assumption might well be favoured by evolution in a culturally transmitted c-induction task. This concludes the analysis of variations on the co-evolutionary model. The following section presents a discussion of the insights these analyses offer.

4.4 Discussion

I classified the preceding analyses broadly into three kinds. Firstly, analyses of: asymmetric fitness differentials between languages (section 4.1.1); fitness benefits from learning rather than communication (section 4.1.2); and continuous genetic types (section 4.1.3), all relate to assumptions about the process of biological evolution. Second, analyses of learning strategies intermediate between MAP and sampling (section 4.2.1), and of populations in which both of those strategies are present (section 4.2.2), along with the analysis of learners who must learn among continuously varying languages (section 4.2.3), all concern the nature of individual language learning. Third, the analysis of learning from multiple teachers (section 4.3.1) concerns a change in assumption about the process of cultural transmission. These themes represent the three core ingredients of the evolutionary process under study in this thesis: cognitive adaptation through co-evolution is built upon reciprocal interactions between biology, learning, and culture. With the variations analysed here, I have tested alternative assumptions in all three fundamental components of the co-evolutionary system. The remainder of this discussion is split into two parts: in section 4.4.1, I discuss whether these variations on the co-evolutionary model have revealed new insights that change the basic conclusions of the previous chapter; section 4.4.2 presents a general discussion of the overall implications of these findings given all analyses.

4.4.1 New Insights from Alternative Assumptions

4.4.1.1 New Support for Nativism?

Which if any of these manipulations has revealed an evolutionary dynamic that supports strong linguistic nativism where the equivalent basic model did not? Just one: the numerical analysis of idealised populations of sample learners who gain increased
fitness benefits by learning language $L_1$ rather than $L_0$ (section 4.1.1.2). In those populations co-evolution delivered strong innate constraints that biased learners to acquire $L_1$ and caused a linguistic universal. However, even this case is questionable: simulations of exactly the same process in finite populations of sample learners (section 4.1.1.3) failed to recover the result, even assuming a strong fitness asymmetry between languages. Though the results of stochastic simulations are less precise than the numerical analysis, they capture stochastic effects that must have influenced human evolution. Formal analysis of finite population mathematical models of this process, which capture stochastic effects but retain analytic precision, would be a valuable extension to the analyses I have undertaken. Notwithstanding this unresolved case, none of the manipulations I have considered provides new support for the evolutionary defence of strong linguistic nativism.

4.4.1.2 New Support for the Amplification Effect?

Have any of the preceding manipulations given any reason to question or support the plausibility of amplification dynamics, and the associated co-evolutionary pattern that leads to weakly biased learners? Where evolution was able to choose between learners whose biases are amplified in cultural transmission and those whose biases aren’t (section 4.2.2), the former reliably out-compete the latter. Where I have tampered with the process of cultural transmission and allowed learning from multiple teachers (section 4.3.1), this has resulted in amplification and the associated co-evolutionary dynamic where before these were absent. This second result rests on a combination of facts about individual psychology and cultural transmission. Where the transmission process permits observation of behaviour in multiple individuals, this creates an opportunity for minor tweaks in individual psychology to drive amplification: in this case, the irrational predisposition to treat all linguistic observations as being representative of a single language. Mechanisms such as these can be extremely valuable, and may be plausible adaptations themselves. In the context of cultural transmission these mechanisms buy a cheap, fast, and robust route to coordination not available to a-cultural species.

This is how culture can solve the problems it creates. For instance, the models I have analysed show that, without amplification, cultural transmission through expression and inductive inference can cause the sub-PG problem and stifle coordinated cognitive
specialisation. Cognitive biases that don’t act to substantively constrain prior beliefs about individual languages, but rather bias the inference procedure such that, when plugged into a cultural system, improve the learners chances of coordinating with peers despite weak substantive preferences, can overcome the sub-PG problem and work magic for the co-evolutionary process. MAP learning achieves this by tampering with an induction strategy such that the learner uses only a subset of the information available in the posterior distribution: it need only know the mode of that distribution. Treating linguistic utterances that were generated by different speakers as if they reflect a single underlying language achieves this by tampering with the learner’s knowledge of data likelihoods. Interestingly, both of these inferential biases involve discarding information, rather than attaining new information or performing more complex computations; both are plausibly cheap cognitive changes. Forgetting, or throwing away some subset of the information available during learning, has also been identified as a necessary condition for coordination in signalling games, by models of cognition and transmission quite different to those I have employed here (Spike et al., 2013).

It is an intriguing possibility that biases of this sort represent a kind of degradation: it raises the possibility that they could emerge through relaxation of selection pressures, or have little to do with selection at all. Arguments against linguistic nativism typically feature claims about the statistical prowess of domain-general inferential capacities (e.g. Saffran et al., 1996). However, the idea that just-off rational inferential biases might aid language acquisition, and drive the evolution of communicatively optimal languages, is not without precedent. For instance, N. Smith et al. (2013) suggest that certain signatures of recursive pragmatic reasoning are well explained by a computational model of pragmatic inference that is constrained to be irrational in very specific ways. One of those constraints corresponds to the assumption that language learners believe in the existence of a single, to-be-learned consensus language, despite learning from independent individuals. This is very similar indeed to the irrational behaviour (based on Smith, 2009)9 that led to amplification in the multiple teacher co-evolutionary models I analysed. These models suggest that mechanisms of this kind can have dramatic consequences for the interaction between biology and culture: as such, the search for mechanisms that lead to amplification in the context of cultural transmission may be of central significance to our understanding of human evolution.

9Note that N. Smith et al. (2013) and Smith (2009) are different Smiths!
Chapter 4 Questioning the Basics: Alternative Co-evolutionary Analyses

4.4.2 General Discussion in Light of All Analyses

4.4.2.1 How specific to language are these conclusions?

The mechanics of the language model reflect arbitrarily distinct mental representations and associated behaviours. The production procedure reflects a standard probabilistic sampling model, and the inference procedure follows generic principles of rational belief updating over a space of competing hypotheses. Though these procedures are thought to be characteristic of language, they are not specific to linguistic behaviour: results may accordingly apply to any behaviour that relies on these cognitive procedures. I have assumed two related fitness functions: one that rewards coordination, and one that rewards accurate learning. Again, these are appropriate assumptions for language, but may also apply to other structured systems of coordinated knowledge that are transmitted culturally through learning from examples, such as music or artistic styles for example.

4.4.2.2 Overall Implications for Nativism

In chapter one, I defined nativism as a tendency to hypothesise strong innate inductive biases as the explanation for skewed typological distributions of linguistic structures, and as a solution to the poverty of the stimulus. Likewise, I defined empiricism as a rejection of these same inductive biases. What is the most general statement of the implications of these results for the argument that evolutionary reasoning supports linguistic nativism on these terms? Wherever a behaviour is culturally transmitted through expression, inference, and induction, and that process of cultural transmission results in amplification of individual cognitive biases, evolutionary reasoning does not support nativism. At best, assuming no costs to specialisation, evolutionary reasoning shows that nativism is as plausible as any theory that appeals to innate biases of any magnitude, but does not support nativism over those other possible theories. If stronger constraints on learning are at all costly or less likely than weaker biases to emerge through mutation, evolutionary reasoning renders nativism implausible. Strong innate constraints are never the optimal solution to the evolutionary problem. Wherever a behaviour is culturally transmitted through expression, inference, and induction, and that process of cultural transmission does not lead to amplification of individual cognitive biases, the evolutionary rationale for nativism about that behaviour is considerably less
convincing than equivalent reasoning about an a-cultural behaviour. In this scenario, the evolution of strong substantive innate constraints on learning is dependent upon favourable assumptions. Of the two kinds of cultural transmission, and the associated co-evolutionary patterns, the first is strongly favoured by selection, and is, to a far greater extent than the second, consistent with empirical facts about our evolutionary history, our individual psychology, and our genetic make-up (see the discussion section of chapter 3).

However, in the previous chapter I questioned the coherence of the conclusion that, in the case of amplification, neither nativism nor empiricism in the strict sense is supported by evolutionary reasoning. The results of these models point to ways in which the definition of these doctrines can be relaxed, such that the core premise of each is maintained, and can be shown to be not only compatible, but also simultaneously supported by evolutionary reasoning that accommodates culture. For example, a weaker version of nativism might be characterised by a tendency to believe that regularities in language typology and acquisition demand that humans are equipped with some degree of domain-specific inductive bias. Likewise, a weaker version of empiricism can be characterised as a tendency to minimise commitment to innate domain-specific inductive biases. The results of these models show that both of these tendencies can at once be valid, if they obey limits suggested by the evolutionary predictions. For the nativist, the intuition that facts about typology and acquisition demand idiosyncratic inductive biases in cognition can be upheld if subjected to the condition that those biases are weak in strength. For the empiricist, the intuition that human cognition cannot be strongly pre-prepared with restrictive biases is also upheld, subject to the condition that cognition is allowed to be just-off neutral; commitment to domain-specific innate inductive biases can be minimised, but cannot be zero. In other words, there is a space where these weaker theses overlap. If (and only if) we take cultural transmission into account, evolutionary reasoning leads precisely to this space. The limits on these theses suggested by the model are fairly unobjectionable to both camps.

Sidestepping the lines of theoretical division momentarily, what do these results say? On the basis of evolutionary reasoning, the default expectation of specialised cognition for culturally transmitted behaviours like language should be weak substantive inductive biases, alongside general inferential biases that lead to amplification; this is the most robust, and most conservative, evolutionary prediction. Conversely, where
strong innate biases that guide language acquisition are discovered, the models I have analysed suggest that these are unlikely to have evolved through natural selection for improved communication. At least, that explanation for their existence should not be automatically assumed plausible on the grounds that similar explanations are plausible for a-cultural behaviours. Adaptive accounts of strong language-specific strong innate inductive biases are not robust.

4.4.2.3 General Implications for Approaches to Language

Despite questioning the evolutionary plausibility of nativism as traditionally conceived, the findings of these analyses might also be construed as reconciliatory. Questions of the evolution of language and cognition often attract polarised debate, in which biological and cultural approaches appear quite incompatible (Christiansen and Kirby, 2003). Seen in the right light, these analyses might offer room for greater integration of perspectives. For the cultural evolutionist, I hope these results show that biological evolutionary explanations need not minimise the role of culture. Rather, those processes can be radically manipulated by culture. Furthermore, cognitive adaptation need not necessarily imply strong or inviolable constraints on cognition. As a result, specialised cognition may be extremely difficult to detect experimentally: consequently, these findings arguably motivate the instantiation of cultural transmission\textsuperscript{10} in any experiment designed to reveal innate language-specific cognitive biases. For those who believe that cognitive specialisation is an indispensable component of linguistic theory, and that adaptation through biological evolution under natural selection for improved communicative abilities is the right explanation for this specialisation: rejoice! These results show that cultural processes aren’t necessarily the enemy of cognitive adaptation. Rather, there is at least one way that cultural evolution can actually make cognitive adaptation considerably faster and more robust. All that must be abandoned is the commitment to strongly constraining innate predispositions, in favour of appeal to weak innate biases. Finally, those who advocate non-adaptationist evolutionary accounts of language often disfavour classic evolutionary reasoning on the grounds that: i) it implies slow and gradual adaptation, yet language emerged rapidly (e.g. Bolhuis et al., 2014); or ii) the value of newly emerged, non-shared specialisations would be too

\textsuperscript{10}Or the use of analyses that make predictions about cultural transmission. These issues will be the subject of chapter 5.
low for them to be maintained and to spread through a population. The weak biases predicted by these co-evolutionary models emerge rapidly, and are sufficiently cheap to plausibly proliferate even if only initially bestowing small fitness benefits. These results show that, like non-adaptationist evolutionary mechanisms, adaptation under natural selection for improved communication too can be rapid, can draw on cognition-external regularities (culture), and thus can bring about big effects from small genetic changes.

In summary, the model variations I have analysed in this chapter support the basic conclusions drawn from the main co-evolutionary model. The major prediction of all these models is that, if and where language-specific innate predispositions exist, they are likely to be weak. Models like these are often criticised on the grounds that their predictions, and therefore their credibility, cannot be evaluated empirically (e.g. Hauser et al., 2014; Watumull and Hauser, 2014). A natural way to empirically test the prediction I have outlined would be to search for weak biases in artificial language learning experiments. In fact, there exists a recently burgeoning enterprise in uncovering linguistic biases with exactly this technique (see e.g. Tily and Jaeger, 2011; Culbertson, 2012). However, the arguments I have developed in this and the previous chapter may seem to spell bad news for this task: if we do have evolved language-specific biases, they’re likely to be so weak that we have no hope of uncovering them experimentally. In the final chapter of this thesis, I show that this need not be the case. I’ll argue that a keen consideration for bias magnitude could distinguish the kinds of biases predicted by these models, and even be a productive turn for experimental approaches to this topic in general. Cognition for culture may be quirky, but there are techniques that allow us to accommodate these quirks, and those techniques might also open up a range of exciting additional opportunities.
Chapter 5

Uncovering Weak Biases Experimentally: Quirks and Opportunities

The co-evolutionary models I have derived and analysed in previous chapters suggest that any domain-specific constraints on cognition for cultural behaviours like language are likely to be weakly biasing. A common criticism of models like these is that their predictions cannot be evaluated empirically (e.g. Hauser et al., 2014; Watumull and Hauser, 2014). To be sure, the models I have derived, and the predictions they support, are necessarily abstract. Worse, it is an explicit prediction that the kinds of biases that could evolve to support cultural behaviours like language may be difficult to detect in individuals. This chapter is a response to those concerns.

With appropriate techniques, weak inductive biases can, if they exist, be uncovered experimentally. I’ll advocate an approach that combines artificial language learning experiments, cognitive model fitting, and techniques for making predictions about cultural transmission. In section 5.1, I’ll briefly review contemporary research in this vein, and consider additional challenges that result from the prediction that domain-specific biases could be weak and defeasible. Through analysis of experimental and simulated datasets, I’ll highlight a potential technical difficulty for this approach and propose a candidate solution (section 5.2). In section 5.3, I’ll suggest that these techniques also unlock opportunities to make predictions about the population-level consequences of weak biases on the basis of experimental data; fitting cognitive models to experimental
data creates opportunities to plug participants into models of cultural evolution.

5.1 Uncovering Cognitive Biases for Language

Characterising aspects of cognition that explain the existence of linguistic universals is a long-standing endeavour. Among linguists, the traditional approach to this task has involved deriving formal linguistic principles that account for attested universals while ruling out unobserved linguistic structures. Such principles then represent hypotheses about constraints on learning imposed by innate, language-specific cognitive machinery. In parallel, the traditional approach within psychology has involved experimental or observational study of first and second language acquisition. However, a third approach to the problem is becoming increasingly popular: the design and analysis of artificial language learning experiments.

5.1.1 Biases in Artificial Language Learning

Artificial language learning experiments have long been used to study biases in language acquisition. Recently, however, there has been an explosion in the use of artificial language learning experiments to directly test hypotheses about the cognitive underpinnings of typologically well-attested linguistic structures. In general, the approach involves constructing a miniature artificial language that contains some linguistic property that is typologically common, training participants on that language, and testing their ability to learn the property of interest. The chief advantage of this approach over comparable typological or formal research lies in the construction of tightly controlled learning problems; experimental evidence that artificial languages with typologically common properties are easier to learn than those with typologically rare properties points directly to a cognitive basis for linguistic universals. Thorough reviews of the approach can be found in Tily and Jaeger (2011) and Culbertson (2012). The method has been applied to universals in various domains of language, including syntax (Smith and Tsimpli, 1995; Christiansen, 2000; Tily et al., 2011a; Culbertson et al., 2012; Culbertson and Adger, 2014), morpho-syntax (St Clair et al., 2009; Hupp et al., 2009; Fedzechkina et al., 2012; Ramscar, 2013), and phonology (Finley and Badecker, 2012; Finley, 2012; Wilson, 2006; Moreton and Pater, 2012a,b; White, 2013).

At least two general classes of experimental paradigm are common. Some studies
(e.g. Christiansen, 2000) involve training one group of participants on an artificial language that contains the typologically attested property, and others on languages that contain rarer alternatives, then testing whether learners of the more common variant have done a better job. Assessments of how well participants have learned include, for example, grammaticality judgements or generalisation to unseen parts of the language. Another approach is also becoming common. Rather than distribute the common and rare linguistic structures between participants, several recent studies (e.g. Hudson Kam and Newport, 2009; Fedzechkina et al., 2012; Culbertson et al., 2012) have employed what Culbertson et al. (2012) call the mixture-shift paradigm. In this case, participants are trained on an artificial language that contains a mixture of the common and rare linguistic forms. At test, participants are required to produce utterances from the language. If, among participants’ productions, the proportion of utterances containing the common variant significantly exceeds their proportion in the training data, this is taken as evidence for a bias in favour of that variant.

Both approaches have their merits (see e.g. Tily and Jaeger, 2011, for discussion), and are already providing new insights. However, in the remainder of this chapter, I’ll concentrate on the mixture shift paradigm, since it is an instance of a broader class of experiments that investigate the origins of regularity in language (to be discussed further in section 5.2) that will prove particularly relevant.

Artificial language learning experiments promise to provide "...a powerful way to detect subtle universal biases on acquisition" (Tily and Jaeger, 2011, pp. 498). However, I have argued that evolutionary accounts for the origins of language-specific cognitive biases will vary considerably in plausibility according to the strength of those biases: for example, adaptationist claims are considerably less plausible when applied to strongly constraining biases. While identifying the presence or absence of linguistic biases is important, it is equally crucial, when trying to understand their nature and origins, that we are able to judge how tightly they constrain cognition: the full picture requires techniques to quantify their strength.

5.1.2 Bias Quantification Through Cognitive Model Fitting

Obtaining estimates of inductive bias through comparing formal models to experimental data has a strong history in contemporary cognitive science (e.g. Haussler, 1988), and is gradually becoming common in artificial language learning studies (e.g. O’Donnell
et al., 2005; Wilson, 2006; Culbertson and Smolensky, 2012; White, 2013). In particular, several efforts have recently been made to obtain quantitative measures of the biases underpinning language by constructing Bayesian cognitive models and directly inferring estimates of their bias parameters from participants’ behaviour in artificial language learning experiments (Reali and Griffiths, 2009; Culbertson, 2012; Ferdinand et al., 2013). This endeavour is particularly relevant to my concerns because, crucially, it allows direct comparison with the kinds of bias predicted by the evolutionary models I have derived: the same kinds of cognitive model are used in both. It is promising indeed that these techniques are available. However, when the object of inquiry is bias strength, and particularly if those biases might be weak, this might entail technical difficulties that would not arise if, for instance, we were only looking for the presence or absence of a cognitive module that imposed inviolable constraints on learning. In the next section, I consider some of these, and work through an example to demonstrate that they can be overcome.

5.2 Model Fitting & Weak Biases: Quirks

I have suggested that a synthesis of artificial language learning experiments and cognitive model fitting techniques provides a viable method for testing the prediction that domain-specific biases are likely to be weak. However, the approach faces clear challenges. For example, the statistical signatures of weak inductive biases might also be relatively weak in experimental data. Detecting these patterns might require unusually many data-points, for instance. There may also be more fundamental difficulties. For example, the potential for weak biases to drive observable linguistic regularities arises partly because, in the context of cultural transmission, learners tend to encounter data that don’t heavily contradict those biases. The power of weak biases derives from the fact that, when learning from other learners, the input to language acquisition has already passed through the minds of similarly biased learners: weak biases are effective because they act over input that has the right properties. Put simply, it might prove difficult to see the effects of biases that are built to be effective in cultural conditions with experiments that don’t mimic those cultural conditions.

One solution to this problem that is gaining in popularity (see e.g. Scott-Phillips and Kirby, 2010) involves experimentally instantiating cultural transmission as a means to observe the population-level consequences of inductive biases, as their effects accumu-
late over generations (e.g. Kalish et al., 2007; Kirby et al., 2008; Griffiths et al., 2008; Reali and Griffiths, 2009; Smith and Wonnacott, 2010). In some scenarios, this methodology alone is sufficient to directly reveal inductive biases by converging on distributions of linguistic forms that faithfully reflect those biases. In other scenarios though (for instance where convergence to the prior cannot be assumed *a priori*) model-fitting techniques are an attractive complementary method for quantifying inductive biases. However, part of the reason that instantiating cultural transmission is so useful is because doing so allows learners, rather than experimenters, to dictate the properties of the training language. This highlights a factor that may be problematic for reverse engineering biases underpinning culturally learned behaviours: the properties of the training language influence the learner’s productions. This is problematic because, in simple terms: we are trying to reverse engineer cognitive biases on the basis of behaviour, but behaviour in these experiments reflects cognitive biases and experience with a training language. Does the nature of the training language influence our ability to detect a cognitive bias?

### 5.2.1 Accounting for Input

Consider the following scenario. We hypothesise the existence of a weak bias in favour of some linguistic property, and wish to expose that bias using an artificial language learning experiment in the mixture shift paradigm. Thus, we must: i) train participants on a language that provides more or less evidence of that property; ii) elicit productions based on their knowledge of that language, and; iii) make inferences about participants’ biases based on whether their productions demonstrate significantly more or less uptake of the property than is justified by the evidence they observed. How should we decide, when designing a training language for step i), on the degree of evidence for the hypothetically favoured linguistic feature? Particularly where inductive bias is weak, the statistical properties of the training language will have considerable influence on the learners’ behaviour. Can this have consequences for our impression of learners’ cognitive biases?

There is at least one case where these consequences may have led to a confusing conclusion. In their Noun-Modifier word-order pattern experiments, Culbertson and Smolensky (2012) (C&S) find evidence for a substantive bias in learning that arbitrarily disfavours one of four possible ordering patterns. This bias is in line with the typo-
logical distribution of these patterns, such that the pattern disfavoured by learning is also typologically rare. C&S go on to quantify the nature of this bias by constructing a Bayesian computational model of word-order learning, and inferring from their experimental data the bias parameters which best account for participants’ behaviour. The results of this model fitting confirm the presence of a bias against the typologically rare structure. However, these clear waters are muddied by the details of the model-fit, which portray a bias that is implausibly strong. Of this inference, C&S say:

"the model assigned a prior weight of 0 to the component of the bias that favoured the typologically rare pattern 4 (Adj-N, N-Num)....the results of our model...may appear to predict in some sense that no language should use deterministic pattern 4. The typological data reveal, however, that 4% of the world’s languages have pattern-4-dominant grammars....this suggests that in a larger sample of experimental data (or a sample with a larger set of training utterances), some learners should be able to acquire such a pattern.”
(Culbertson and Smolensky, 2012, pp. 1492)

Reformulating, whilst the typological distribution and C&S’s experimental results do indicate a bias against the rare pattern, it is clear that humans are able to learn it, as is seemingly exemplified by the children born into communities in which are spoken those 4% of the world’s languages that do employ the otherwise rare pattern. Why then do C&S’s sophisticated bias-inference techniques suggest their participants are subject to a hard constraint that prohibits acquisition of the rare pattern?

There may of course be many explanations for this inconsistency, including those highlighted in the paper (too few participants and/or too few training utterances). However, it is worth noting the possibility that the inconsistency is at least in part down to the particular set of training conditions deployed in these experiments. Whilst the possible ordering patterns are well-balanced across C&S’s training regimes, the possible proportions of those patterns in the training mixtures are sparsely represented. Since I have suggested above that learning behaviour (upon which quantifiable bias inferences are drawn) is influenced by the kinds of utterances learners observe, it may be that this particular mixture proportion encourages behaviour that, when reverse engineered through model fitting, portrays a bias stronger than that under which it was in fact generated; or, put another way, other training proportions may be more likely to induce behaviour that more accurately portrays the true learning bias. To illustrate this point trivially, we might hypothesise that learners exposed to a training language
that contained only the rare pattern (a 100% : 0% mixture), as are children who learn natural languages that exclusively or nearly-exclusively employ this pattern, would produce learning behaviour that portrays a weaker bias. Combined with data from existing training regimes, these data might dilute the conclusions of C&S’s model-fitting such that its predictions better suit typological observations.

In C&S’s case, the potential inconsistency is clear: the inference is unusually strong, and typological data are sufficient to cast doubt on its accuracy. Furthermore, while plugging these gaps in the experimental data might change our understanding of the kind of bias against this word-order pattern that learners possess, it is unlikely to change the direction of our understanding: the case for the existence of some bias in this domain seems strong. However, there may be cases where the idiosyncrasies of our data aren’t so clear. And there may be cases where those idiosyncrasies may have more dramatic consequences for our understanding. In the next section, I’ll explore these issues by fitting a cognitive model to data from Smith and Wonnacott’s (2010) artificial language learning experiment, in which participants were trained on a range of input languages.

5.2.2 Example 1: Smith and Wonnacott (2010)

5.2.2.1 Experimental Details

Smith and Wonnacott (2010) (S&W) carried out artificial language learning experiments in order to explore the emergence of regularity in linguistic systems. Languages universally tend to avoid unpredictable variation: “In general, no two linguistic forms will occur in precisely the same environments and perform precisely the same functions.” (Smith and Wonnacott, 2010, pp. 444). Much research is dedicated to the possibility that this regularity is brought about by a regularisation bias that favours the acquisition of regular linguistic systems (e.g. Kam and Newport, 2005; Hudson Kam and Newport, 2009; Smith and Wonnacott, 2010; Reali and Griffiths, 2009; Perfors, 2012; Ferdinand et al., 2013). Several authors have argued that this bias may be domain-specific for language (e.g. Reali and Griffiths, 2009; Ferdinand et al., 2013), and that it can be well characterised as an inductive bias (e.g. Reali and Griffiths, 2009; Perfors, 2012). The regularisation bias is a particularly useful test case for at least two reasons: first, it has been suggested that the bias may be weak (e.g. Reali and Griffiths, 2009; Smith and Wonnacott, 2010); second, several authors have explored regularisation through the use of
Bayesian cognitive models (e.g. Reali and Griffiths, 2009; Perfors, 2012; Ferdinand et al., 2013).

S&W explored the emergence of regularity in plural marking systems. Participants\(^1\) each learned an artificial language that specifies how to mark plurality for four cartoon animals, by observing sentences paired with visual scenes. For each animal, for each presented sentence, plurality was marked by the presence of one of two possible plural markers (post-nominal morphs). Each participant observed sentences demonstrating how to mark plurality (for each of the four animals) 12 times (12 naming events). Crucially, the level of regularity in those plural marking systems was determined by variation in the use of the two plural markers between naming events. For example, maximally regular systems would employ just one of the two markers to indicate plurality \textit{in all 12 naming events} for a given animal. In contrast, minimally regular (maximally variable) systems would employ both plural markers in equal proportion for each animal. Thus, training languages can exhibit more or less regularity dependent on the relative proportions in which the two plural markers were employed.

After learning these systems (training), participants were asked to describe similar scenes (testing), thus producing a set of utterances in which plurality is marked with one of the two markers. By comparing the regularity of plural marking systems exhibited in productions with those the participant had observed in training, a measure of regularisation behaviour is obtained. S&W instantiate cultural transmission using single-teacher diffusion chains. Specifically, 10 chains of five generations each were instantiated. The first learner in each chain learned a language whose regularity was specified by S&W: for those participants\(^2\), plurality was marked for each animal by a "majority marker" that appeared in 75% of relevant naming events, alongside the "minority marker" that appeared in the remaining 25% of relevant naming events. For the remaining learners in the chains, their training language was formed from the productions of the learner at the previous generation. What makes this dataset so useful is that, because training languages were determined by cultural transmission rather than fixed at some specified level of regularity experimentally, participants observed a variety of marker proportions. In the following section, I’ll describe a Bayesian cognitive

\(^{1}\)Sixty five monolingual English-speaking undergraduate Psychology students. Here, I’ll only summarise the details of these experiments directly relevant to my analyses. For full experimental details, see Smith and Wonnacott (2010)

\(^{2}\)And 15 “isolate” learners.
model for this learning task, fit the model to S&W’s data, and ask how our impression of learners’ biases varies according to the training language they observed.

5.2.2.2 A Model for Smith and Wonnacott’s Data

Though many other models might be equally suitable, I’ll adopt a Bayesian probabilistic model based on the Beta and Binomial distributions, for two reasons: first, this exact model has been used to capture the regularisation bias in other regularisation studies (e.g. Reali and Griffiths, 2009; Ferdinand et al., 2013), and is very similar to those used in other treatments of regularisation experiments (e.g. Perfors, 2012) and for cognitive biases underpinning language more generally (e.g. Culbertson and Smolensky, 2012); second, it is very similar to the model of language I assumed in the main co-evolutionary model, and in fact is almost identical to the model of infinite PCFG learning employed in chapter 4.

Assume each participant saw a dataset \( y^{t-1} = (y^{t-1}_1, y^{t-1}_2, y^{t-1}_3, y^{t-1}_4) \). Each \( y^{t-1}_j \) is a count of the number of times animal \( j \) appeared with the "majority" plural marker, so \( y^{t-1}_j \in \{0, 1, ..., n^{t-1}\} \), where \( n^{t-1} = 12 \) is the total number of naming events per animal for all animals, and it follows that the count for the "minority" marker for animal \( j \) is \( n^{t-1} - y^{t-1}_j \). On the basis of these observations, the learner attempts to infer the underlying true probability of each plural marker. Since there are only two, and these probabilities must sum to 1, we can simply keep track of the probability of the majority marker, \( p_j \in [0, 1] \), and note that the probability of the ‘minority’ marker for animal \( j \) is \( 1 - p_j \). The inferential task is to compute a posterior distribution \( P(p_j|y^{t-1}_j) \) over the possible hypotheses \( p_j \). I assume the learner does this by combining the likelihood that it would observe these data if \( p_j \) were the true probability, \( P(y^{t-1}_j|p_j) \), with a prior probability, \( P(p_j) \) of the hypothesised \( p_j \), in accordance with the principles of rational inference, using Bayes’ rule so that:

\[
P(p_j|y^{t-1}_j) = \frac{P(y^{t-1}_j|p_j)P(p_j)}{P(y^{t-1}_j)}. \tag{5.1}
\]

Here the numerator \( P(y^{t-1}_j) \) is a normalising factor giving the probability of the ob-
served data averaged over all hypotheses, and can be largely ignored for these purposes. The data likelihoods are given by the Binomial sampling scheme, so that:

$$P(y^t_j - 1|p^t_j, n^t-1) = \binom{n^t-1}{y^t_j} p^t_j^{y^t_j} (1 - p^t_j)^{n^t-1 - y^t_j},$$  \hspace{1cm} (5.2)$$

and the prior probability of $p_j$ follows a Beta distribution with matched shape parameters $\alpha$, so that:

$$P(p_j) = \frac{p^\alpha_j - 1(1 - p^\alpha_j)^{\alpha-1}}{B(\alpha, \alpha)},$$ \hspace{1cm} (5.3)$$

where $B(x, y)$ denotes the standard beta function. The Beta distribution is the conjugate prior for the $p_j$ parameter of the Binomial sampling scheme, meaning that the posterior distribution takes the same familial form as the prior, affording various mathematical conveniences. These distributions are widely used in combination, and have been widely used to capture human learning (e.g. Reali and Griffiths, 2009; Chater et al., 2010; Culbertson and Smolensky, 2012). The parameter $\alpha$ characterises the inductive bias of the learner, with differing values privileging different regions of the $[0,1]$ domain. Where $\alpha > 1$, middling values of $p_j$ are favoured (i.e. values closer to 0.5), representing a preference for variability. Where $\alpha < 1$, extreme values of $p_j$ (close to 0 or 1) are favoured, consistent with a bias toward regularity. In statistical notation, the model is:

$$p_j \sim \text{Beta}(\alpha)$$

$$y^t_j \sim \text{Binomial}(p_j)$$

Given this learning model, it is possible to search for the value of $\alpha$ that, given the training data participants have seen, predicts their responses at test with the highest probability. To do this, I assume that, for each animal in its dataset, the learner induces an estimate $\hat{p}_j$ of the true underlying $p_j$ by drawing a random sample from the posterior distribution, so that $\hat{p}_j \sim P(p_j|y_j, \alpha)$. On the basis of this estimate, the learner generates utterances that are distributed according to the Binomial sampling scheme described above, such that $y^t|n^t \sim \text{Binomial}(\hat{p}_j)$, where $y^t$ is the count of the number of times the learner produced the ”majority” marker for animal $j$ at test, and $n^t$ is the total number
of utterances elicited for that animal during testing for that participant. Of course, we cannot see the estimate \( \hat{p}_j \) actually induced by the participant, so we must integrate out this uncertainty, by calculating, for all possible values of \( p_j \), the probability that the learner would have induced that estimate, and the probability of the observed testing productions if it had, as in (5.4). Thanks to the conjugacy of the Beta and Binomial distributions, it can be shown that:

\[
P(y_t^j|y_{t-1}^j, \alpha) = \int_0^1 P(y_t^j|p_j)P(p_j|y_{t-1}^j, \alpha) \, dp_j
\]

\[
= \int_0^1 P(y_t^j|p_j)P(y_{t-1}^j|p_j)P(p_j|\alpha) \, dp_j
\]

\[
= \binom{n_t}{y_t^j} \frac{\text{B}(\alpha + y_{t-1}^j, \alpha + n_{t-1} - y_{t-1}^j)}{\text{B}(\alpha + y_t^j, \alpha + n_t - y_t^j)}.
\]

It is also straightforward to calculate this quantity under the alternative assumption that learners induce \( \hat{p}_j \) by selecting the maximum a posteriori (MAP) hypothesis: the value \( p_j \) that maximises \( P(p_j|y_{t-1}^j, \alpha) \). Again there exists a convenient expression for the MAP hypothesis in this model 5:

\[
p_{j}^{\text{max}} = \frac{\alpha + y_{t-1}^j}{2\alpha + n_{t-1}},
\]

where \( p_{j}^{\text{max}} \) is the MAP hypothesis. If \( \hat{p}_j = p_{j}^{\text{max}} \), then eq. (5.4) reduces to the binomial sampling probability of the productions given that estimate, so that:

\[
P(y_t^j|y_{t-1}^j, \alpha) = \binom{n_t}{y_t^j} \left( p_{j}^{\text{max}} \right)^{y_t^j} \left( 1 - p_{j}^{\text{max}} \right)^{n_t - y_t^j}
\]

\[
= \binom{n_t}{y_t^j} \left( \frac{\alpha + y_{t-1}^j}{2\alpha + n_{t-1}} \right)^{y_t^j} \left( 1 - \frac{\alpha + y_{t-1}^j}{2\alpha + n_{t-1}} \right)^{n_t - y_t^j}
\]

With these in hand, it is possible to compute, for any set of participants, the value \( \hat{\alpha} \) that maximises eq. (5.4) (or eq. (5.6)). In practice, I’ll report the value \( \hat{\alpha} \) that maximises the logarithm of the product of eq. (5.4) or eq. (5.6) (the combined log-likelihood).

\[\text{Note that I am using the posterior mean, not the mode, as the MAP estimate, following Reali and Griffiths (2009)}\]
for each datapoint in a given set of participant responses, and the geometric mean of the likelihood as the average per-datapoint predictive power of that estimate (See Lewandowsky and Farrell, 2011).

The first task is to fit the model to the full dataset of IL chains (10 chains of 5 generations each) and isolates (15), including all learning episodes (4 animals per participant) for all participants (65), totalling 260 learning episodes. For this dataset, under the sampling hypothesis selection assumption, the maximum likelihood\(^6\) bias estimate is \(\hat{\alpha} = 0.314\), which assigns the data a log-likelihood of -279.303, predicting the outcome of each learning episode with average probability 0.342. This indicates a medium strength bias towards regularity. Under the MAP hypothesis selection assumption \(\hat{\alpha} = 0.562\), indicating a weaker bias for regularity. The MAP model assigns the data a lower log-likelihood of -314.347, predicting each datapoint with average probability 0.298. Given the superior predictive power of the sampling assumption, I will continue with this model during the remaining calculations for this dataset.

Recall that S&W’s data are particularly useful for my purposes here because they include data from learners that were trained on a range of possible input marker proportions. Participants observed, for each of the four animals, one of five possible input marker proportions: \(\frac{0}{4}, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, \frac{4}{4}\). Since \(n^{t-1} = 12\) across the board, observed majority marker counts\(^7\) for all animals were in practice constrained so that \(y^{t-1}_j \in \{0, 3, 6, 9, 12\}\).

The iterated learning chain experimental method means that, other than for those learners who constituted the first generations of the various IL chains, and participants in the "isolate" condition, the input proportions on which learners were actually trained were not specified in advance\(^8\), but instead determined, for each participant, by the behaviour of the participant at the preceding generation of her IL chain. This means the full dataset may include more episodes of learning from some input proportions than from others. Given that I am exploring the potential influence of differing training regimes on bias estimation, this is not to be overlooked. Ignoring all other factors (generation, participant, animal, etc.), and simply counting up the number of learning episodes from each possible input proportion, we have \(|\frac{0}{4}| = 74, |\frac{1}{4}| = 64, |\frac{2}{4}| = 12, ...

\(^6\)All function optimization procedures were achieved computationally with appropriate optimization algorithms from the Python package Scipy.

\(^7\)I will sometimes refer to these possible majority marker training proportions as input regimes, for brevity.

\(^8\)Though these input proportions were constrained to take certain values by experimental procedure, as described above.
\[ \left| \frac{3}{4} \right| = 60, \text{ and } \left| \frac{1}{4} \right| = 50, \text{ where } \left| \frac{i}{4} \right| \text{ indicates the number of learning instances during which a learner observed the majority marker used in proportion } \frac{i}{4}. \text{ This makes clear, for instance, that many more learners observed maximally regular utterances (majority marker counts of 0 or 12) than observed maximally variable utterances (majority marker counts of 6). This is to be expected, since in these experiments the diffusion chains converged on regular languages relatively quickly, meaning a large proportion of participants in later generations encountered regular utterances. Has this asymmetry in input regimes influenced our impression of learners’ biases?; Do certain input regimes reveal certain biases better than others?}

To answer these questions, we can interrogate the dataset to ask: if we had only seen data from some subset of the possible training regimes, what would be our impression of the regularisation bias? In the most straightforward case, if we fit the model to data exclusively from specific training proportions, will the estimate of \( \alpha \) change at all? Figure 5.1 shows the results of fitting the model to these data. It displays maximum likelihood values for \( \hat{\alpha} \) by input marker proportion.

As figure 5.1 shows, constraining our data to single input regimes appears to cause notable differences in the inferences we would draw about the nature of participants’ regularisation biases. Had participants only been trained on the \( \frac{1}{4} \) or \( \frac{2}{4} \) input regimes, we would have inferred the existence of a hard constraint: the strongest possible regularisation bias; had we only trained participants on the \( \frac{3}{4} \) input regime, we would conclude from their behaviour that they likely possess a very weak regularisation bias. Regimes \( \frac{0}{4} \) and \( \frac{4}{4} \) would indicate reasonably strong and medium strength regularisation biases respectively. Though this picture is concerning, there may be some qualifying factors. In particular, as noted previously, these regime ‘bins’ contain different amounts of data: the \( \frac{2}{4} \) condition, for instance, contains only 12 learning episodes, while \( \frac{0}{4} \) contains 74. It may be then that the regimes with more data are giving a more accurate picture of the true bias, while our estimates based on the sparser regimes are simply noisy: this is one of the explanations Culbertson & Smolensky point to for their implausible model-fit, and seems a sensible possibility here too, since (notwithstanding any content bias for one or the other marker types, which we have no independent reason to suspect) complementary regimes \( \left( \frac{0}{4} \text{ & } \frac{1}{4}; \frac{1}{4} \text{ & } \frac{3}{4} \right) \) should lead to comparable results, but don’t.
This issue does not prevent a fairer approach to inference of a bias from the full dataset. Though it will be a noisier estimate, I can make efforts to ensure each training regime contributes equally to our inference. This requires model inference over a dataset in which each training regime contains an equal number of datapoints. That number is constrained to be no greater than 12, since the smallest training regime in the dataset, $\frac{2}{14}$, contains only 12 unique datapoints. Rather than simply throw away arbitrarily chosen subsets of the regimes with more data, it is possible to repeatedly re-sample size-12 random subsets of data for regimes that contain more than 12 datapoints, fit the model to each of these sampled datasets, and report summary statistics for $\hat{\alpha}$. Figure 5.2 shows the results of fitting $\hat{\alpha}$ to 10000 such equally-balanced samples of the data. The mean of these $\hat{\alpha}$ values is $0.355 \pm 0.026$; $0.003$, representing a slightly weaker bias than inferred from the full dataset. The histogram shows a skew towards lower values, as we should expect when, for example, the influence of regime $\frac{2}{14}$, which individually suggested the strongest possible regularisation bias (figure 5.1), is increased.
Though S&W's data have been useful to illustrate these issues, there are aspects of the dataset that make it imperfect for the analysis: namely the asymmetry inherent in S&W's data just discussed, and potential shortfalls of the model in capturing all aspects of learner behaviour in this task\textsuperscript{9}. To navigate these problems, we can explore these issues on an idealised dataset, where all factors can be controlled, by simulating the ideal experiment.

5.2.3 Example 2: A Simulated Experiment

With the frequency-learning mode described above in hand, it is possible to generate learning data in exactly the conditions favourable to this analysis. The advantage of this approach is that I am specifying in advance the true learning bias from which this data will be generated. On this basis, we can then imagine that this bias were unknown, perform the inferential analysis, and precisely judge the quality of our conclusions against the true generative model.

\textsuperscript{9}In this experiment, learners may have drawn inferences about the marker proportions for one animal conditional on their beliefs about markers for other animals. The model does not account for this possibility.
Chapter 5  Uncovering Weak Biases Experimentally: Quirks and Opportunities

First, I will simulate the following scenario. While I’ll continue to use S&W’s topic and experimental procedure as illustrative, the same dynamics may apply to frequency learning in many comparable tasks. Imagine learners in S&W’s experiment observed just one animal being named \( n^{t-1} = 12 \) times. Each time the animal is observed being named, it appears with one of the two possible plural markers, exactly as in S&W. On the basis of these observations, learners are asked to produce another \( n^t = 12 \) utterances describing the scene (i.e. to produce twelve instances of plural marking for that animal themselves). Roughly following S&W’s framework, I will simulate an experiment in which 30 participants learn from each of the possible input proportion regimes \( \frac{0}{1}, \frac{1}{1}, \frac{2}{1}, \frac{3}{1}, \frac{4}{1} \) (i.e. 30 participants per regime = 150 participants total). For the productions of these simulated participants from each regime, I will compute a maximum likelihood estimate \( \hat{\alpha} \) of the learning bias just as before. To smooth over the effects of random noise in the experiment, I will simulate this experiment a large number (10000) of times, and report for each training regime the summary statistics of estimates \( \hat{\alpha} \) averaged over simulated experiments. Given that I must specify a bias to be shared by all participants, I will repeat this analysis twice, once having specified a strong-ish regularisation bias \( \alpha = 0.1 \), and again with a moderate bias towards variable systems \( \alpha = 5 \). Figure 5.3 shows the results of model fits to the first simulated dataset, in which a strong regularisation bias \( \alpha = 0.1 \) was specified.
Figure 5.3: The logarithms of the mean values for $\hat{\alpha}$ by input regime over 10000 simulated frequency learning experiments in which simulated learners have a strong bias towards regularity. Errorbars give the logarithms of 95% confidence intervals in the raw $\hat{\alpha}$ samples. The dashed horizontal line illustrates the true $\alpha$. The solid horizontal line at 0 indicates the root of each bar, and the unbiased prior. $n^t = n^{t-1} = 12$; 30 participants per input regime per experiment, $\alpha = 0.1$ ($\ln(\alpha = 0.1) = -2.3$).

Note that in this case, the figure shows $\ln(\hat{\alpha})$ rather than $\hat{\alpha}$, for two reasons: first, in $\hat{\alpha}$, differences across some input regimes are large, making it difficult to display some results at a useful comparative resolution, whereas in $\ln(\hat{\alpha})$ these differences are smaller and can be presented on the same scale without loss of clarity; second, the log transform has the nice property that $\ln(\hat{\alpha}) < 0$ represents a bias towards regularity, $\ln(\hat{\alpha}) > 0$ represents a bias towards variability, and $\ln(\hat{\alpha}) = 0$ represents the uniform prior, or the absence of bias. In fact, figure 5.3 shows the logarithm of the mean of the raw $\hat{\alpha}$ samples, $\ln(E(\hat{\alpha}))$, for each input regime. Also displayed are the logarithms of the 95% confidence intervals (again, computed in the raw -not log transformed- $\hat{\alpha}$ samples) for $\hat{\alpha}$ for each input regime. The solid horizontal line at 0 at once indicates the root of the
bars (so that, for example, a bar that descends from zero indicates a negative $\ln(E(\hat{\alpha}))$ at its lowest point) and the unbiased prior ($\ln(\alpha) = 0$) value. The dashed horizontal line at -2.3 indicates the logarithm of the true $\alpha = 0.1$ bias specified beforehand.

It is clear from these results that, at least for this magnitude of data, and for this bias, different input regimes have systematically influenced the outcome of the model fit. The fully regular regimes ($\frac{0}{12}$ and $\frac{12}{12}$) lead to behaviour from which the model-fit has on average accurately identified the generating bias, and consistently identified the presence of a regularisation bias. Fits to data from the moderately variable regimes ($\frac{3}{12}$ and $\frac{9}{12}$) also lead to inference of a regularisation bias, but consistently under-estimate its strength. Inferences based on data from the maximally variable regime ($\frac{6}{12}$) appear a great deal more varied than those based upon data from other regimes, and consistently lead to the false conclusion that learners in this experiment are likely to possess a bias in favour of variable systems, often a very strong bias in this direction. Though confidence intervals for the maximally regular and moderately variable/regular regimes overlap, Welch’s t-tests confirm significant differences in the expected outcome of fitting models to data gathered in these regimes ($\frac{0,12}{12}$ & $\frac{3,9}{12}$, $t \approx \pm 51$, $p < 0.001$). Expected outcomes of complementary regimes ($\frac{0,12}{12}$ & $\frac{12,12}{12}$, $\frac{3,9}{12}$ & $\frac{9,12}{12}$) naturally show no significant differences under the same tests. Expected outcomes of the maximally variable regime are significantly different to all other regimes ($\frac{0,3,9,12}{12}$, $t \approx \pm 7.95$, $p < 0.001$).

Comparable outcomes across complementary regimes, and meaningful differences between distinct regimes, indicate the influence of a systematic bias\(^{10}\) in this procedure, rather than noisy parameter estimates, at least in the case of a strong regularisation bias.

Figure 5.4, which is directly comparable to figure 5.3, shows results of an equivalent analysis of simulated experiments in which a weak bias favouring variability ($\alpha = 5$) was specified in advance. In this case, all except the maximally variable regime do a good job of identifying the correct bias on average. Inferences based on data from regime $\frac{6}{12}$ again exhibit notable variability: though here these inferences identify the correct direction of bias (favouring variability), our estimates of the nature of this bias again tend to inflate its strength ($\hat{\alpha} > \alpha$).

---

\(^{10}\)A bias in the inference procedure, not a learning bias
Figure 5.4: The logarithms of the mean values for \( \hat{\alpha} \) by input regime over 10000 simulated frequency learning experiments in which simulated learners have a weak bias towards variability. Error-bars give the logarithms of 95% confidence intervals in the raw \( \hat{\alpha} \) samples. The dashed horizontal line illustrates the true \( \alpha \). The solid horizontal line at 0 indicates the root of each bar, and the unbiased prior. \( n^t = n^{t-1} = 12; \) 30 participants per input regime per experiment, \( \alpha = 5. (\ln(\alpha) = 0.693)\)

5.2.4 Input Regime Honesty

The examples above illustrate that, at least for some biases, some input regimes might skew our inferences about those biases. Why? One way to think about the problem is to consider what I will term the *Honesty* of a particular input regime for a particular bias or set of biases. That is, given a learner with a particular bias, how faithfully, on average, will that bias be reflected in the data the learner produces if she has been trained in an input regime \( i \)?

It is possible to formulate this question mathematically, and at least take steps towards an understanding. Intuitively, we wish to calculate for a given input regime the
following. For all possible types of utterance a learner could produce at test, what is the likelihood that, given her bias and her training regime, she would produce each utterance type, and given each utterance type and our model fitting procedure, how accurate would our estimate of her bias be? By computing a measure of our inferential accuracy for a given utterance type, weighting that measure by the likelihood of the utterance in these conditions, and combining these quantities for all utterances, we have a measure of the Honesty of that input regime for that bias.

Formally, continuing with the same model and notation as above, let’s say that, for an input regime (majority marker count) \( i \in \{0, 1, \ldots, n^{t-1}\} \), and a particular bias \( \alpha \), the Honesty of that regime for a learner with that bias, \( H^i_\alpha \), is defined as follows. The desiderata for this measure are loosely these: the quantity should lie on some interpretable range; it should measure how accurate the maximum likelihood inference is; and it should take into account our confidence in the maximum likelihood estimate. First we must know how likely a particular outcome or set of productions is: for notational clarity, when referring to the likelihood of some production given some bias \( \alpha \) and some input regime \( i \), I will use the following shorthand\(^{11}\).

\[
L^\alpha_{ij} \equiv P(y^t = j \mid y^{t-1} = i, \alpha),
\]

(5.7)

where \( P(y^t = j \mid y^{t-1} = i, \alpha) \) is as defined in (5.4). Next, we require a measure, \( \rho(\hat{\alpha}, \alpha) \), for the accuracy of our inference: a simple place to start is to measure the absolute difference between the inferred bias and the true bias, so that \( \rho(\hat{\alpha}, \alpha) = |\hat{\alpha} - \alpha| \). Where precisely the correct bias is inferred, this quantity will be zero; where a different bias is inferred, \( \rho(\hat{\alpha}, \alpha) \) can be arbitrarily large, since in this particular model the range for \( \alpha \) is unbounded. Finally, to quantify how strongly the model-fit supports the inferred bias over the true bias, the Honesty measure will include the ratio of likelihoods between the inferred and the true biases. This quantity will be 1 where \( \alpha = \hat{\alpha} \), and approach zero as the likelihood surface becomes more peaked at \( \hat{\alpha} \neq \alpha \).

So, ranging over possible outputs, the Honesty of a particular input regime for a particular bias is defined to be:

\(^{11}\)Note that I am now using lowercase letter \( j \) to refer to possible marker counts rather than particular ‘animals’ as before.
\[ H_i^\alpha = \sum_{j=0}^{n^i} \frac{L_{ij}^\alpha}{L_{ij}^\alpha_j} \cdot \frac{1}{1 + \rho(\alpha, \hat{\alpha}_j)} \cdot L_{ij}^\alpha, \]  

(5.8)

where \( \hat{\alpha}_j \) is the maximum likelihood bias inferred from productions with majority marker count \( j \). Though the expression can be simplified, I have left separate the terms to the right of the sum to make clear the contribution of each of the factors discussed above. The final term represents the likelihood of producing the majority marker \( j \) times in these conditions; the middle term represents the difference between the inferred bias and the true bias, and the leftmost term gives the ratio of likelihoods for outcome \( j \) given the inferred and true biases. Figure 5.5 shows \( H_i^\alpha \) for the input conditions tested in the simulated experiments above, both for the case of a pre-specified regularisation bias \( (\alpha = 0.1) \) and a bias favouring variability \( (\alpha = 2.0) \). These measures naturally tally with the empirical findings shown above in figures 5.3 and 5.4: for both biases, maximally variable input is least Honest; for a strong regularisation bias (red bars in figure 5.5), more regular input languages are more Honest; for a weak bias favouring variability (blue bars in figure 5.5), moderately variable input regimes are most Honest.  

![Figure 5.5: Honesty of various input regimes for \( \alpha = 0.1 \) (red) and \( \alpha = 2 \) (blue).](image)

A clear objection to the utility of this measure is that, to compute the Honesty of
a particular training regime, the bias must be known in advance. If the object of the experiment is to uncover an unknown bias, then the measure is little help. Against this charge I note the following: first, it may in some cases be true that for all possible biases, some input regimes are more/less honest than others. The analysis above suggests that this might be true, for instance, of the maximally variable regime in the model under consideration: for at least two distinct bias kinds, this regime is less Honest than others. In such cases our uncertainty about the true bias does not prevent us from selecting an optimal set of training regimes by first computing their Honesty.

Second, it may often be the case that we do in fact have a priori reason to suspect that participants’ true bias lies in some subset of the possible range of biases. The regularisation bias is a clear example of such foreknowledge: experimental evidence, among other types, strongly suggests the presence of some bias favouring regular linguistic structures. The open question, which is central to the predictions made in this thesis, concerns the nature of this bias: is it a hard constraint or a defeasible predisposition?; can we determine its strength?; is it stronger or weaker in the linguistic domain than in other domains? In this case we can bound our uncertainty about $\alpha$ to the set of regularisation biases - to $\alpha \in (0, 1)$ - and account for this bounded uncertainty when computing the Honesty of possible training regimes. To do this requires only that we integrate over the range during Honesty calculations. If there is reason to believe a bias $\alpha$ lays in some range $\bar{\alpha} = [a, b]$, the Honesty of input regimes for biases in that range can be obtained as in (5.9):

$$H^\alpha_{\bar{\alpha}} \sim \int_a^b H^\alpha_\alpha \, d\alpha$$  \hspace{1cm} (5.9)

Figure 5.6 shows the Honesty of all possible input regimes (for learners observing $n^{t-1} = 12$ and producing $n^t = 12$ naming events) averaged over the possible range of regularisation biases (over 100 points between $\alpha = 0.001$ and $\alpha = 0.99$). Though the differences in Honesty between input regimes are softer than those observed for the strong regularisation bias alone ($\alpha = 0.1$), the trend is preserved: less variable $=$ more honest for this range of biases. Note that the Honesty of input regimes in figure 5.6 is generally lower than the Honesty of equivalent input regimes for the strong regularisation bias shown in figure 5.5. This is because the range of biases covered by figure 5.6 includes the weaker regularisation biases: at least in this case, Honesty is lower across the board for weaker biases.
The calculations I have undertaken here are specific to this learning problem and this model. However, the idea that we can interrogate a learning model to calculate Honesty a priori may be widely applicable. An Honesty measure gives us at least one method to gain some foreknowledge about the consequences of the training languages we construct. It allows us to answer questions such as: which regimes will in theory facilitate more accurate bias inference?; and which regimes are likely to give rise to conclusions that would generalise well to other training conditions?

An interesting hypothesis, which I leave open for future research, is that iterated learning chains naturally converge toward the more Honest input regimes. Smith and Wonnacott’s data are consistent with this hypothesis: their iterated learning chains converged relatively quickly on maximally regular input languages, which are, according to the calculations above, the most Honest input regimes for revealing a regularisation bias. If the hypothesis were correct, iterated learning chains would provide a useful
method to identify the optimal conditions for revealing inductive biases in scenarios where it is difficult to calculate Honesty in advance, for instance. It would also motivate the experimental instantiation of cultural transmission as a means to uncover biases even if convergence to the prior cannot be assumed.

In summary: inferences about the nature of cognitive biases based on data gathered in artificial language learning experiments may be systematically influenced by the languages on which participants have been trained; some input regimes may be more Honest than others. This may be particularly important for detecting weaker biases. I have considered some straightforward techniques that can illuminate these differences a priori and help to guide the choices that must be made during experimental design and model fitting. Given the complexities I have outlined, is fitting cognitive models to experimental data worth the hassle? I argued above that, if we are interested in the nature of constraints on cognition underlying cultural behaviours, it is. In the next section, I’ll briefly consider another motivation. If weak biases can have large effects on cultural systems, then to understand their consequences we must be able to make predictions about the outcome of cultural transmission. Crucially, the techniques that help us uncover weak biases also present new opportunities to analyse their long-term effects.

5.3 Model Fitting & Weak Biases: Opportunities

The evolutionary arguments I have made about weak biases depend upon the idea that they can, over the course of cultural transmission, give rise to dis-proportionate population-level effects. To whatever extent it is sensible to search for weak biases, it is also necessary to acknowledge that we can only understand their relationship to universals by considering how their effects accumulate over generations of transmission. It is worth noting that this factor has received comparatively little attention in the literature relating artificial language learning results to linguistic universals. Though the promise of methods that account for cultural transmission has been acknowledged (Tily and Jaeger, 2011), it is still common practice to make claims about universals, which arise over many generations of transmission, on the basis of "single generation" experiments (e.g. Culbertson et al., 2012; Culbertson and Adger, 2014).

One direct way to plug this gap is to instantiate cultural transmission experimen-
tally (Scott-Phillips and Kirby, 2010). However, we may wish to make predictions about cultural processes that are difficult, expensive, or impossible to instantiate in the lab, such as long term transmission or large populations. The scope of experimental instantiations of culture is restricted by resources. Conversely, scope is the chief strength of formal analyses of cultural transmission: they can illuminate processes too complex or large-scale to be easily studied experimentally. For instance, language evolution research has been transformed by the use of mathematical and computational models of cultural transmission to illuminate processes previously un-studyable (see Steels, 2011, for a review). Of course, the chief weakness of formal analyses of culture is their potential lack of connection to human psychology. Artificial language learning experiments and formal analyses of culture have complementary strengths and weaknesses.

In general, there is very little research concerning the potential to make predictions about cultural processes based on data from experiments that don’t actually instantiate those cultural processes. One exciting method is the construction and analysis of empirical transition matrices (see Ferdinand et al., 2013, for details): these side-step the requirement for (and potential complications of) a cognitive model, and directly extrapolate from learning data to long-term transmission, but rely on quite specific experimental conditions. Here I concentrate on another technique. By fitting cognitive models to data from artificial language learning experiments, and using those models as the basis for formal analyses of cultural transmission, we can combine the strengths of both approaches: it allows us to plug participants into models of culture. In the following sections, I consider a few examples of how this might be achieved. Though there are many ways to make predictions about cultural processes, I consider two in particular: in section 5.3.1 I explore how cognitive model fitting can facilitate the construction and analysis of transition matrices; in section 5.3.2 I suggest that cognitive models that have been tuned to humans can also form the basis of computational simulations.

5.3.1 Using Models to Construct Transition Matrices

Cultural transmission can be understood as a stochastic process with the Markov property (Griffiths and Kalish, 2007). One way to study the dynamics of cultural transmission is to define the transition matrix associated with this process. Once we have de-

---

12 Though on-line experiments using crowd-sourcing platforms hold promise in this direction (See e.g. Tily and Jaeger, 2011)
fined a transition matrix, standard methods of matrix analysis can then be used to gain numerous insights into the behaviour of the process, such as its stationary distribution, or the rate at which it will converge on particular states. Though there are various ways to go about this task, here I’ll simply illustrate how cognitive model fitting can be used to construct a transition matrix from participant behaviour, continuing with the model and learning problem of the preceding sections.

In this example, a transition matrix is a stochastic matrix \( Q \) whose components \( q_{i,j} \) give the probability that a participant who has learned a training language with majority marker count \( i \) will produce that marker \( j \) times during testing. The transition matrix therefore defines how probable are the transitions between all possible input marker counts and production marker counts. The probability of each individual transition can be computed from the cognitive model: crucially, this probability is influenced by the bias, \( \alpha \), which can be fit to experimental data as described earlier. The following short sections illustrate this procedure for different kinds of experimental data.

5.3.1.1 One Prior Fits All

In the simplest case the procedure is as follows: choose a model of learning; infer for that model the best fit parameters from the entire experimental dataset; compute the transition matrix under that model; conduct analyses of that matrix for the quantities of interest. In this case, then, entries in the matrix are determined straightforwardly by the posterior predictive distribution given the bias that has been inferred, so that:

\[
q_{i,j} = L_{ij}^{\hat{\alpha}} = P(y^t = j | y^{t-1} = i, \hat{\alpha})
\]  

(5.10)

where \( \hat{\alpha} \) is the maximum likelihood estimate of a single prior based on all participants’ data. This approach could be appropriate when all participants behave similarly and a single model fits all participants’ behaviour well. Likewise, it could be appropriate in situations where we have a well balanced, or particularly Honest, set of input proportions represented in our data. Reali and Griffiths (2009) follow exactly this procedure in their regularisation experiments, and show that a weak regularisation bias can have strong effects on an artificial language over generations.
5.3.1.2 To Each his Own Prior

In some cases a single model fit may be inappropriate on the basis that participants exhibit considerable individual differences, or a range of biases exist among participants. Where no single model predicts all participants’ behaviour well, and there is a reasonable magnitude of data from each participant, a sensible approach is to fit model parameters to each participant’s behaviour individually. Essentially, this approach allows us to create a transition matrix that describes the dynamics of cultural transmission among a well mixed population of individuals who may have different biases. In this case, assuming a population of \( N \) participants in an experiment, a transition matrix can be treated as a mixture of \( N \) equally probable models, so that:

\[
q_{i,j} = \frac{1}{N} \sum_{k=1}^{N} L_{ij}^{\hat{\alpha}_k} \tag{5.11}
\]

where \( \hat{\alpha}_k \) is the bias inferred from participant \( k \)’s data. It may seem counterintuitive to simply average over participant models when computing transition probabilities. The logic behind this calculation is clearer when considering a population of \( k \) distinct learners, each of whom may be selected as teacher or learner with uniform probability in a single-teacher iterated learning chain.

5.3.1.3 Hierarchical Models & Distributions of Priors

In many cases participants’ biases may be subject to individual differences reflecting genuine variation that can be quantified and accounted for in these calculations. For instance, assume a set of \( N \) individual best-fit biases \( \bar{\alpha} = \{ \hat{\alpha}_1, \ldots, \hat{\alpha}_N \} \) has already been inferred, one for each participant, such that \( \hat{\alpha}_k \in \bar{\alpha} \) is the bias inferred from participant \( k \)’s data. Variation among these biases may be genuine, and the size \( N \) sample may be a sparse realisation of this variation.

To smooth over these sparse observations, we can assume they are drawn from some distribution, and infer the parameters of that distribution that best predict the observations in much the same fashion as I have laid out for participant responses. With this we are adding an extra layer of hierarchy to our model in order to account for uncertainty about the true distribution of biases in the population of interest. For example, in the regularisation experiments as discussed above, we might have inferred a set of individual biases \( \bar{\alpha} \) that contains regularisation biases of varying strengths, many clus-
tering around some central value, with fewer inferred biases further from that central value. We might assume that these biases can be treated as independent samples from a population whose variation can be captured by a Gaussian distribution with unknown parameters \( \bar{\pi} = \{ \mu, \sigma \} \). By calculating the values \( \hat{\pi} \) that maximise \( P(\alpha | \bar{\pi}) \), we obtain an estimate of the shape of this distribution, or the nature of the variation present in the population of biases. When computing the transition matrix, we can use this distribution as a probability measure over the range of possible biases and integrate over the range, weighting the individual contributions of possible biases accordingly, as in eq. (5.12):

\[
q_{i,j} \sim \int_{\alpha} L_{ij}^\alpha P(\alpha | \hat{\pi}) \, d\alpha
\]  

(5.12)

where in this example \( P(\alpha | \hat{\pi}) \sim N(\alpha; \hat{\pi}) \). There is in principle no reason why, in cases where variation among participants is crucial to an analysis, we cannot go further and integrate out our uncertainty about \( \bar{\pi} \). That is, rather than assume confidence in the maximum likelihood parameter estimates \( \hat{\pi} \) and hold the distribution of biases in the shape determined by those values, it is possible to integrate over the possible shapes of this distribution (the possible values of \( \hat{\pi} \)), as in eq. (5.13):

\[
q_{i,j} \sim \int_{\alpha} \int_{\hat{\pi}} L_{ij}^\alpha P(\alpha | \hat{\pi}) P(\hat{\pi} | \bar{\alpha}) \, d\alpha \, d\hat{\pi}.
\]  

(5.13)

Where the properties of experimental data permit, the computation in eq. (5.13) can be made more manageable by choosing a conjugate prior \( P(\bar{\pi}) \) for \( \bar{\pi} \) which provides an analytic expression for the posterior predictive distribution \( P(\alpha | \bar{\alpha}, \beta) \), where \( \beta \) is a hyperparameter governing \( P(\bar{\pi}) \), which captures uncertainty in \( \hat{\pi} \) without the need for heavy numerical integration. In this case eq. (5.13) becomes:

\[
q_{i,j} \sim \int_{\alpha} L_{ij}^\alpha P(\alpha | \bar{\alpha}, \beta) \, d\alpha.
\]  

(5.14)

The example above utilising a Gaussian distribution to capture variation among biases, for instance, would allow various such convenient choices of prior. There is also no need to restrict such an analysis to single parametric distributions. Though I’ll skip the technical details here, it would be straightforwardly possible to infer a variety of distributions of priors for use in transition matrix calculation, for cases where multiple separable learning strategies are present among participants.
Though the merits of these methods for making long-term predictions will vary by case, I hope to have demonstrated here that, in principle, difficult datasets need not prohibit analyses of this kind. Markov chain analysis can be a powerful method for illuminating cultural transmission: cognitive model fitting opens up the use of these methods to make predictions about cultural transmission even when our experimental data are insufficient to define a transition matrix empirically. In theory, wherever there exist appropriate cognitive models, this technique can even be used retrospectively on data from existing experiments that were not designed with culture in mind. Whether or not these techniques can make reliable predictions is, of course, an empirical question: benchmarking their accuracy against experimental instantiations of the same process may be a widely valuable endeavour. Figure 5.7 gives a simple schema of the methods I have outlined.
A. Empirical Transition Matrix

B. Single Prior Transition Matrix

C. Individual Priors Transition Matrix

D. Distribution of Priors Transition Matrix

Figure 5.7: A schematic presenting a range of methods for computing transition matrices from experimental data.  
A: Populating a transition matrix with empirically observed (input, output) datapoints.  
B: Computing a transition matrix from a model whose parameters have been fit to a complete dataset.  
C: Populating a transition matrix by fitting model parameters to individual participants, computing transition matrices for each, and averaging over/mixing those individual matrices to obtain the transition matrix corresponding to a well-mixed population containing these biases.  
D: Computing a transition matrix by fitting model parameters to individual participant data, inferring the distribution(s) of those parameters/biases in the participant population, integrating over the range of possible biases, mixing the transition matrices defined by each, while weighting their contributions according the inferred distribution.
5.3.2 Computational Simulations

Finally, it is also worth noting that, once a cognitive model has been fit to experimental data, cultural transmission under that model can be explored through computational simulations. This method has a huge history in the study of the cultural evolution of language (see e.g. Steels, 2011), for instance, and affords incredible flexibility. While transition matrix analysis is powerful and useful for some cultural processes, they cannot easily capture complex transmission dynamics. For example: transmission in populations that split into sub-populations (e.g. Boyd and Richerson, 1985; Baronchelli et al., 2012); transmission among populations that are spatially structured (e.g. Kirby, 2002); learning from multiple teachers (e.g. Burkett and Griffiths, 2010; Niyogi and Berwick, 2009); and late introduction of new biases (Gialluisi et al., 2013) are all examples of transmission dynamics that might be difficult to capture with transition matrix analysis, but could be straightforwardly simulated computationally.

In this chapter, I have addressed some of the empirical consequences of the prediction that any language-specific cognitive biases are likely to be weakly constraining. I have advocated a synthesis of artificial language learning experiments, cognitive model fitting, and formal analyses of cultural transmission. I have argued, with the support of several examples, that this approach simultaneously offers the potential to discover weak cognitive biases and to make predictions about their population-level consequences. I have considered several potential difficulties with the approach, and suggested solutions.
Conclusion

In this thesis, I asked whether the human capacity to transmit knowledge culturally could have influenced the evolution of our cognitive apparatus. To explore this question, I derived and analysed a series of mathematical models that capture the general dynamics of the co-evolutionary process brought about by cultural transmission. I argued that these models cast doubt on the two prevailing views of cognitive adaptation to cultural behaviours: that it is directly comparable with adaptation to a-cultural behaviours (e.g. Pinker and Bloom, 1990; Pinker and Jackendoff, 2009b); or that it is outright implausible (e.g. Christiansen and Chater, 2008; Chater et al., 2009).

According to the models I have explored, cultural transmission is a game-changer for the evolution of cognition: in several ways, culture can intervene in the evolution of hard-wired knowledge. The evolutionary rationale for nativism should be reconsidered with respect to culturally transmitted behaviours. However, while it complicates the evolutionary process, culture can solve the problems it creates: where innate dispositions can be amplified over the course of cultural transmission, small cognitive changes can give rise to disproportionately large consequences. In this way, culture buys a cheap, robust route to specialisation; it can speed up cognitive adaptation, allowing biology to keep up with culture through minor tweaks. I argued that this mode of evolution has considerable explanatory potential: it resolves the apparent paradox in the fact that we observe cognitively driven population-level universals but extreme individual-level plasticity; it explains why domain-specific innate biases may be difficult to detect in individuals, but potentially observable in large-scale population-level analyses; it explains why the human genome shows signs of increased, recent, rapid adaptation concomitant with the onset of culture; and it explains how coordinated cognitive specialisation could, in the case of language, have emerged so quickly.

Above all, the evolutionary perspective I have outlined in this thesis makes a simple
and general prediction: if and where evolution has built coordinated cognitive biases to support a behaviour that is culturally transmitted through inductive inference, we should expect those biases to be weak and defeasible. I have argued that this prediction can be tested, and suggested techniques that could help us to do so. Ultimately, of course, the nature of the biases that underpin cultural behaviours should be determined empirically. An understanding of the processes that shape cognition can be a powerful guide for that endeavour. The findings conveyed in this thesis represent a small step towards that understanding: they suggest that culture can play a profound role in the evolution of cognition, but that we are only just beginning to understand how; they are, after all, just maths.
References


233


236


241


243


