

BIASES AND STAGES IN PHONOLOGICAL ACQUISITION

A Dissertation Presented

by

ANNE-MICHELLE TESSIER

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2007

Linguistics

© Copyright by Anne-Michelle Tessier 2007

All Rights Reserved

BIASES AND STAGES IN PHONOLOGICAL ACQUISITION

A Dissertation Presented

by

ANNE-MICHELLE TESSIER

Approved as to style and content by:

Joseph V. Pater, Chair

John J. McCarthy, Member

Lyn Frazier, Member

Shelley L. Velleman, Member

Elizabeth O. Selkirk, Department Head
Linguistics

DEDICATION

To Elizabeth Tessier and Attilio Favro.

ACKNOWLEDGEMENTS

It has now become cliché to begin one's acknowledgements by acknowledging that they are the part most likely to be read in any linguistics dissertation. I apologize for the cliché but not for its truth; since I plan to continue reading other people's acknowledgements before all else, I can only offer the hope that readers enjoy these acknowledgements enough to read on until perhaps the table of contents.

First I want to express my gratitude to my committee members: Joe Pater, John McCarthy, Lyn Frazier and Shelley Velleman. My chair Joe Pater has guided me gently through every obstacle of graduate school; I thank him for helping me construct and then improve every piece of this dissertation and for always making the effort to figure out what I was trying to think or write and extracting the salvageable parts. I also owe him for involving me in his own work from the very beginning, and for teaching me about being a theoretician who runs experiments, a phonologist who talks to psychologists, and a linguist who still goes to rock concerts. I thank John McCarthy for being, well, John McCarthy: for applying the full power of his empirical and technical knowledge to all of my work, for reading and improving drafts nearly before I gave them to him, and for being so devastatingly clear. I also thank him for teaching the first UMass phonology seminar I attended, which convinced me I was doing the right thing, and for setting the bar for teaching phonology and Optimality Theory so instructively high. I thank Lyn Frazier for being a voice of reason, breadth and perspective, for her experimental knowledge, for always asking the trickiest questions, and for sharing her enthusiasm for connecting data and theory, making strong claims and thinking very hard. I thank Shelley

Velleman for providing her expertise on phonological acquisition and disorders, for carefully understanding my perspective and carefully reading my work, for pointing me to all the data I didn't know, and for smiling in the face of stress.

Beyond my committee, I thank John Kingston for teaching me most of the correct things that I know about doing and thinking about phonetics, and for always being willing to discuss any theory or experiment as clearly as I would let him, and I thank Lisa Selkirk for challenges and support throughout my time at UMass. Among UMass phonology students past and present, I thank Kathryn Flack, Michael Becker, Shigeto Kawahara, Tim Beechey, Matt Wolf, Maria Gouskova, Andries Coetzee, Mike Key, Karen Jesney, Kathryn Pruitt, Della Chambless, Jonah Katz and Dan Mash, for data, theory, advice, argumentation, proofreading, scripts, stats, commiseration, deep-fried food and all manner of collegial help. I am so very grateful to have worked in this department's unique, talented and supportive community of teachers and peers; I hereby acknowledge how much I owe you all.

Many other UMass people must also be thanked for getting me through grad school. I would like to thank Marcin Morzycki for everything, only I don't know how, so instead I will just thank him for arguing with me – about linguistics, pedagogy and nearly all else, at every hour of the day or night, to extents no doubt in violation of aspects of the Geneva convention – for lending me platypuses, and for making me laugh. With similar insufficiency, I thank Ana Arregui and Paula Menéndez-Benito, for every kind of love, support and sage advice that friends can give, for calling and emailing and checking and comforting and listening, and for making me laugh. I thank Jan Anderssen and Meredith Landman for being such wonderful roommates and sharing so many of the daily horrors

of graduate work and dissertation writing, as well as the food in the fridge, and for making me laugh. I thank my classmates Helen Majewski, Helen Stickney and Shai Cohen, for support both quiet and loud, mutual miseries and reality checks, and for making me laugh. And for many acts of friendship along the way, I thank Keir Moulton, Michael Brigham, Florian Schwartz, Ilaria Frana, Aynat Rubenstein, Kyle Rawlins, Amy Rose Deal, Masako Hirotani and Tessa Warren, as well as everyone else who ever came to Semantics Reading Group and drank my martinis, or made me laugh. I also thank Rajesh Bhatt for dissertation-stimulating music and party-hosting, Angelika Kratzer for being so sweetly happy for me when I got a job, Ellen Woolford for not giving up on the hope that I would someday say something clearly, and Kathy Adamczyk and Tom Maxfield for continually saving my administrative hide.

Outside UMass, I thank the following people whose time, insight, questions and critiques improved many parts of this dissertation: Bruce Hayes, Bruce Tesar, Elan Dresher, Keren Rice, Kie Zuraw, Colin Wilson, Adam Albright, Glyne Piggott, Heather Goad, Marie-Hélène Côte, Kevin Ryan, Jason Riggle, Elliott Moreton, Jennifer Smith, Michael Wagner, and audiences at the University of Toronto, the MOT, HUMDRUM, and Brown/UMass phonology workshops, the 80th LSA meeting, BUCLD30 and WCCFL25. I also thank Alan Prince for his LSA 2005 summer institute course, the Social Sciences and Humanities Research Council of Canada (a.k.a. Money Canada) for their generous support over the last two years, and the Department of Linguistics at the University of Alberta for giving me such a good reason to finish.

At McGill University, I would like to thank my teachers and colleagues who convinced me that I should be a linguist by treating me like one – including Lisa Travis,

Jonathan Bobaljik, Yvan Rose, Evan Mellander, Lydia White, Lara Riente, Charles Boberg and Ben Adaman. In particular I thank Heather Goad, for getting me hooked on phonology, OT and acquisition, sending me to UMass for brainwashing, and then setting an impeccable example of how to support and happily disagree with a former student.

I thank my parents Roger and Rosemary Tessier for always believing I could do whatever I thought I wanted to, for covering all my computational and laundry needs respectively, for their stellar genetic material, and for making me laugh. And I thank Allon Beck and Rebecca Rosenblum for being permanent friends, and for each coming to visit western Massachusetts when friendship at a distance was not enough.

Finally, I thank Kyle Johnson – for teaching me many valuable lessons not discussed in the graduate school handbook, for being charming even beyond the many times he picked up the cheque, for making me laugh, and for preventing me from quitting linguistics when I was at my lowest by pointing out I'd be just as terrible at anything else.

ABSTRACT

BIASES AND STAGES IN PHONOLOGICAL ACQUISITION

FEBRUARY 2007

ANNE-MICHELLE TESSIER, B.A., MCGILL UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Joseph V. Pater

This dissertation presents Error-Selective Learning, an error-driven model of phonological acquisition in Optimality Theory which is both *restrictive* and *gradual*. Together these two properties provide a model that can derive many attested intermediate stages in phonological development, and yet also explains how learners eventually converge on the target grammar.

Error-Selective Learning is restrictive because its ranking algorithm is a version of Biased Constraint Demotion (BCD: Prince and Tesar, 2004). BCD learners store their errors in a table called the Support, and use ranking biases to build the most restrictive ranking compatible with their Support. The version of BCD adopted here has three such biases: (i) one for high-ranking Markedness (Smolensky 1996) (ii) one for high-ranking OO-Faith constraints (McCarthy 1998; Hayes 2004); and (iii) one for ranking specific IO-Faith constraints above general ones (Smith 2000; Hayes 2004).

Error-Selective Learning is gradual because it uses a novel mechanism for introducing errors into the Support. As errors are made they are not immediately used to learn new rankings, but rather stored temporarily in an Error Cache. Learning via BCD is only triggered once some constraint has caused too many errors to be ignored. Once

learning is triggered, the learner chooses one *best* error in the Cache to add to the Support – an error that will cause minimal changes to the current grammar.

The first main chapter synthesizes the existing arguments for this BCD algorithm, and emphasizes the necessity of the Support's stored errors. The subsequent chapter presents Error-Selective Learning, using cross-linguistic examples of attested intermediate stages that can be accounted for in this approach. The third chapter compares ESL to a well-known alternative, the Gradual Learning Algorithm (GLA: Boersma, 1997; Boersma and Hayes, 2001), and argues that the GLA is overall not well-suited to learning restrictively because it does not store its errors, and because it cannot reason from errors to rankings as does the BCD. The final chapter presents an artificial language learning experiment, designed to test for high-ranking OO-faith in children's grammar, whose results are consistent with the biases and stages of Error-Selective Learning.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS.....	v
ABSTRACT.....	ix
CHAPTER	
I INTRODUCTION.....	1
1. Summary of the dissertation	1
2. Structure of the dissertation	3
II PHONOTACTIC LEARNING AND BIASED CONSTRAINT DEMOTION.....	6
1. Introduction.....	6
1.1 Important aspects of Optimality Theory for this learning theory.....	7
1.2 Outline of the chapter.....	8
2. Learning an Optimality-Theoretic Grammar	9
2.1 The learning framework: the Tesar/Smolensky learner.....	9
2.2 The importance of errors and the Support.....	13
2.2.1 The Support and the lexicon.....	14
2.2.2 The Support and the Gradual Learning Algorithm.....	15
3. Restrictive phonotactic learning: Biased Constraint Demotion	16
3.1 Illustrating the BCD approach: high-ranking Markedness	19
3.2 A second bias in BCD terms: high-ranking OO-Faith	23
3.2.1 OO faith as an OT account of cyclicity	24
3.2.2 OO-faith as an OT account of MSCs	26
3.2.3 The learnability argument for high-ranking OO-faith: McCarthy (1998).....	28
3.3 Connecting M >> F and OO >> IO as surface-oriented biases.....	29
3.4 Interim Summary	33
4. An input-oriented ranking bias in BCD: Specific-F >> General-F	33
4.1 The theory of positional faithfulness	34

4.1.1 Why not (only) positional markedness.....	35
4.1.2 Stringency, not fixed rankings.....	42
4.2 The learnability argument for a Specific-F >> General-F bias: Smith (2000).....	43
4.3 The problems of enforcing the Specific-F >> General-F bias.....	46
4.3.1 Language-specific relations between faithfulness constraints.....	46
4.3.1.1 Prince and Tesar's example	46
4.3.1.2 A morphological example	48
4.4 Interim Summary.....	49
5. Returning to the role of the Support	50
5.1 A kind of learning error: winner misparses	51
5.2 How the Support allows BCD to overcome winner misparses.....	53
5.3 A second example	55
5.4 Summary.....	57
6. The proposal: finding the most specific IO-Faith constraint.....	58
6.1 The goal: determining subset relations between the contexts of faith.....	58
6.1.1 Constraint stringency vs. context specificity.....	59
6.1.2 Outline of the proposal.....	60
6.2 The first step: finding universal specificity relations	63
6.3 The second step: finding contingent specificity relations	66
6.4 Why context tables are dynamic	70
6.4.1 What can go wrong in a context table?	71
6.4.2 Overcoming extra marks in a context table	74
6.5 Why contingent specificity cannot be learned from Ls and Ws	75
7. Implementing the Spec-F >> Gen-F bias	77
7.1 A working BCD algorithm	77
7.2 Ways of calculating Spec-F >> Gen-F relations: the Azba case study....	79
7.2.1 Using a context-based F-specificity bias	81
7.2.2 Using constraint-based F-stringency bias: Hayes' simulation	83
7.2.3 Prince and Tesar on the Azba language	85
7.2.4 Summarizing the Azba results	86

7.3	Returning to Anti-Paninian rankings and learning	87
7.4	Summary and outstanding issues	90
8.	Chapter 2 Summary, in preparation for Chapter 3	91
III	ERROR-SELECTIVE LEARNING.....	94
1.	Introduction.....	94
1.1	The approach to reconciling BCD and <i>gradual</i> learning	95
1.2	The Specific Markedness stage: English coda clusters	98
1.3	The Specific Faithfulness stage: French onset clusters	99
1.4	Analytic assumptions about the intermediate stages.....	101
1.4.1	Stringency relations among markedness constraints.....	101
1.4.2	Positional faithfulness and input prosodic structure.....	102
1.5	Roadmap to the chapter.....	105
2.	The data from intermediate stages	105
2.1	Introduction to the data	105
2.1.1	The Compton/Streeter database	106
2.2	Intermediate stages that rely on specific markedness	107
2.2.1	More on complex codas in Germanic.....	107
2.2.2	Markedness of complex onsets, and sonority distance	108
2.3	Intermediate stages that rely on specific faithfulness	115
2.3.1	More on faithfulness in stressed syllables	115
2.3.2	Faithfulness <i>to</i> stressed syllables.....	120
2.3.3	Faithfulness to initial syllables.....	121
2.3.4	Faithfulness to morphological roots	122
2.4	Summary of the data	125
3.	The theory of intermediate stages: Error-Selective BCD	126
3.1	The Error-Selective Learning proposal	126
3.1.2	What happens when an error is made: a Specific-M example....	127
3.1.2	How learning is triggered	128
3.1.3	Step 1: Choosing an error to learn from.....	129
3.1.4	Step 2: Applying BCD.....	130
3.1.5	A second example: a Specific-F stage.....	131

3.2	Discussion of the ESA, and Error-Selective Learning more generally	133
3.2.1	Analyzing the three ESA criteria for choosing errors.....	133
3.2.2	Terminating ESL and converging on the end stage grammar.....	135
3.2.3	Irrelevant markedness violations	137
3.2.4	Choosing among positional faithfulness contexts.....	138
3.2.5	The Violation Threshold and extra-grammatical factors.....	139
3.3	Illustrating ESL: A case study of Trevor and Julia's onset clusters.....	140
3.3.1	Trevor	140
3.3.2	Julia	147
3.3.3	Summary	151
4.	The roles of frequency	152
4.1	The connection between frequency and Error-Selective Learning	152
4.2	The connection between frequency and order of acquisition	153
4.2.1	Data from cross-linguistic frequency: initial weak syllables vs. codas	154
4.2.2	Ambient not output frequencies, and the Error Cache.....	156
4.3	Intermediate stages without stringency: stages of prosodic truncation ..	160
4.3.1	Noting a stringent alternative	165
4.4	Infrequent mistakes and the value of the Error Cache	166
5.	Developmental variation and Error Selective Learning	168
5.1	The ubiquity and challenges of variation in learning	169
5.1.1	The potential for a variable BCD learner.....	170
5.2	Alternative I: the Variable VT approach.....	172
5.2.1	The example of variable codas.....	173
5.2.2	The effects of the variable VT approach.....	177
5.2.3	Deriving developmental regression in the Variable VT approach.....	179
5.2.4	Weaknesses of the Variable VT approach.....	180
5.3	Alternative II: the Cloned Support approach	180

5.3.1	Returning to the variable coda example.....	181
5.3.2	Discussion of the Cloned Support approach.....	184
5.3.3	Regression in the Cloned Support approach.....	185
5.4	Summarizing the variable ESL discussion.....	186
6.	Chapter Summary	187
IV.	THE GRADUAL LEARNING ALGORITHM ALTERNATIVE.....	188
1.	An introduction to the Gradual Learning Algorithm (GLA).....	188
1.1	The GLA view of constraint rankings.....	188
1.2	How the GLA learns a grammar.....	191
1.2.1	The (limited) power of an error in the GLA	193
1.3	Goals and core properties of the GLA	196
1.4	Ranking Biases and the GLA.....	197
1.5	Chapter Roadmap.....	198
2.	Restrictiveness and specific-to-general faithfulness relations in the GLA.....	198
2.1	The exemplifying grammar	199
2.2	The GLA's learning input	200
2.3	The stages of GLA learning	203
2.3.1	The initial state	203
2.3.2	The intermediate stages	204
2.3.3	The end state grammar	206
2.4	Summarizing the results	207
2.4.1	The superset grammar: mid vowels	207
2.4.2	The Anti-Paninian Ranking: round vowels	208
3.	Intermediate stages and the Specific-F >> General-F bias in the GLA.....	209
3.1	The Specific F stages that require the ranking bias	209
3.2	The Specific F stages that don't require the bias: Curtin and Zuraw (2001)	211
3.3	Interim Summary	214

4.	Persistent biases, contingent biases, and the GLA	214
5.	A first problem with not storing errors: winner misparses.....	217
5.1	The GLA's treatment of misparsed winners	218
5.2	Winner misparses and markedness: the same problem	223
6.	Exceptions and end-state variation.....	225
6.1	The GLA's treatment of exceptionality.....	226
6.1.1	Two languages and their codas.....	227
6.1.2	Learning the variable coda grammar.....	228
6.1.3	Learning the exceptional coda grammar.....	230
6.2	Learning exceptions: GLA-related approaches.....	232
6.3	Learning variation without the GLA: a BCD approach.....	236
6.4	Summary.....	243
7.	Chapter Summary.....	243
V.	TESTING FOR THE HIGH-RANKING OO-FAITH BIAS.....	245
1.	Introduction to the chapter	245
2.	The OO-faith ranking bias and phonotactic learning	246
2.1	The role of OO-faith in enforcing restrictiveness	246
2.2.	Predictions for stages of acquisition	249
2.2.1	The target: an OO-unfaithful language	249
2.2.2	OO-faith kicks in at the initial state	250
2.2.3	OO-faith kicks in at an intermediate state	251
3.	The experimental methodology: artificial language learning	253
3.1.	The difficulties in testing for OO-faith in L1 acquisition	253
3.2.	The artificial language learning paradigm	254
3.3	The present application	255
4.	Experimental Design	257
4.1	Experimental predictions	257
4.2	Materials	258
4.3	Methodology	259

4.3.1	Participants	260
4.3.2	Training	260
4.3.3	Testing	262
5.	Experimental Results	263
5.1	The data reported	263
5.2	Testing the predictions	264
5.2.1	Testing prediction 1	264
5.2.2	Testing prediction 2	266
5.3	Summary of the results	267
6.	Rankings in the results	267
6.1	A3's cluster voicing: an intermediate ranking	268
6.2	A3's treatment of coda affricates	269
6.3	N's treatment of [mf.d] clusters: an initial state ranking	271
6.4	Summary of analyses	273
7.	Theoretical discussion	273
7.1	The intermediate stage, and Error-Selective Learning.....	273
7.2	Independent evidence of the OO-faith bias	275
7.3	The persistent OO-faith bias, and the GLA	280
7.3.1	The empirical need for a persistent OO-faith bias	280
7.3.2	The GLA problem with persistent biases and OO-faith	281
8.	Experimental discussion	284
8.1	The connection between natural and artificial language learning	284
8.2	A potential perceptual confound, and the next step	286
9.	Chapter Summary.....	287
	BIBLIOGRAPHY.....	288

CHAPTER I INTRODUCTION

1. Summary of the dissertation

This dissertation presents Error-Selective Learning, an error-driven model of phonological acquisition in Optimality Theory which is both *restrictive* and *gradual*. It is restrictive in that it chooses grammars that can generate observed outputs but as few others as possible; it is gradual in that it requires numerous errors of the same kind to learn a new grammar. Together these two properties provide a model that can derive many observed intermediate stages in phonological development, while still explaining how learners eventually converge on the target grammar.

Error-Selective Learning is restrictive because its ranking algorithm is a version of Biased Constraint Demotion (BCD: Prince and Tesar, 2004), in which learners are biased to prefer rankings between classes of constraints, e.g. Markedness >> Faithfulness. BCD learners store the errors made by their current grammars in a table called the Support, and use their biases to choose the most restrictive ranking compatible with the Support. To account for a range of restrictiveness problems, the proposed learner uses a BCD algorithm with three ranking biases: (i) the well known Markedness >> Faith bias (Smolensky, 1996); (ii) a bias for high-ranking paradigm uniformity constraints (i.e. OO-Faith: Benua, 1997; McCarthy 1998); and (iii) a bias for ranking more specific IO-Faith constraints above more general ones (Smith, 2000; Hayes, 2004).

Error-Selective Learning is gradual because it uses a novel mechanism for introducing errors into the Support. As errors are made they are not immediately used to

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

learn new rankings, but rather stored temporarily in an Error Cache. Learning via BCD is only triggered once some constraint overcomes a Violation Threshold: that is, when some constraint has caused too many errors to be ignored. Once learning is triggered, the error-selective learner assesses the violation profiles of the errors currently in the Cache, and chooses the *best* error to add to the Support – an error that will cause minimal changes to the current grammar. Once the Support has been updated, the error-selective learner uses BCD to build a new ranking, empties their Error Cache and begins again.

In using Violation Thresholds and the proposed method of choosing best errors, the error-selective learner is sensitive to frequency in a way that accords with the connection between order of acquisition and input frequencies in child-directed speech. This reliance on frequency is also used to propose an extension of ESL which derives variation between intermediate stages, by introducing a stochastic notion of Violation Threshold.

Evidence for the Error-Selective BCD model comes from a wide survey of data in the literature, including new results from corpus data, as well as a novel artificial language experiment with children. Error-selective analyses are provided for several intermediate stages, relying on the M >> F and Spec-F >> Gen-F biases, including a detailed examination of the onset cluster acquisition of two children in the Compter/Streeter database. The artificial language study provides novel evidence of the third bias – for high-ranking OO-Faith – by showing that four year old children’s repairs of unfamiliar coda-onset clusters in an ‘alien’ language were skewed in ways that kept derived plural words similar to their singular bases.

At the heart of all the dissertation's proposals is the Support, a stored repository of the data from which the BCD learner has built its current grammar. Through continual updates and revisions to the Support, the BCD learner remains restrictive even in the face of wrong structural analyses and missing data. The error-selective learner is therefore unlike the Gradual Learning Algorithm (Boersma, 1997 *et seq.*), which has until now been the only OT learning algorithm to model gradual phonological learning. It is shown that the GLA is not well-suited to ongoing restrictive learning, even when equipped with a similar series of ranking biases, principally because it does not store its errors.

2. Structure of the dissertation

The first two chapters of the dissertation present the error-selective BCD learning model. Chapter 1 introduces the OT learning approach of Tesar and Smolensky (2000), the problems of restrictiveness in phonotactic learning, and the Biased Constraint Demotion solution. It synthesizes much recent work on restrictiveness in OT learnability theory, and focuses in detail on the implementation of the Specific >> General IO-Faith bias. A thorough method for discovering specific-to-general IO-faith relations is proposed, which compares the *contexts* of faithfulness constraints on a dynamic language-specific basis and uses them to impose the F-specificity bias.

Chapter 2 moves on to the facts of intermediate stages of phonological acquisition, which BCD algorithms on their own are not designed to model. It presents a body of evidence illustrating two kinds of intermediate stages, which both fall between initial and target grammars in their tolerance of marked structures, and then presents the Error-Selective model that derives these stages. It then spells out the role of frequency in

error-selective learning, as embodied in the Error Cache and its constraint violations, and demonstrates the connection between violation frequency and order of acquisition using cross-linguistic data from Germanic, Romance and English (e.g. Roark and Demuth, 2000.) Error frequencies and the Error Cache are then also used to propose a variable version of Error-Selective Learning.

Chapter 3 compares the error-selective BCD learner with an alternative, the stochastic OT learner that uses the Gradual Learning Algorithm (Boersma, 1997; Boersma and Hayes, 2001; Curtin and Zuraw, 2001.) Here it is shown through OTSoft simulations that the GLA must be augmented with all the same biases assumed above: both to learn a restrictive final grammar, and to pass through a full range of attested intermediate stages. Furthermore, the GLA learner can still be tricked into learning superset grammars if the learner makes incorrect assumptions about the learning data. The GLA is also demonstrated to fall short in learning restrictive grammars with lexical exceptions; the current state of learning with regard to exceptionality vs. free variation in the two models remains a central question for further research.

Chapter 4 returns to the notion of ranking biases in learning, using a novel experimental paradigm for artificial language research with children. The data from this experiment, which used a wug-test (Berko, 1958) with both novel morphological bases and a novel suffix, provides evidence that young learners bring a bias for paradigm uniformity to the task of learning novel phonotactic patterns. More generally, these positive results suggest that young children are both willing and able to participate in artificial language learning, pointing to a new source of data in the study of phonological acquisition.

CHAPTER II

PHONOTACTIC LEARNING AND BIASED CONSTRAINT DEMOTION

1. Introduction

The goal of this chapter is to lay out the arguments for a particular view of Optimality-Theoretic learning, instantiated in a class of learning algorithms called Biased Constraint Demotion by Prince and Tesar (2004), henceforth also known as BCD. BCD is principally a model of phonotactic learning – that is, the learner’s discovery of which marked structures are allowed in its language, in what contexts and under what circumstances.

Much of the chapter is a synthesis of the recent relevant learnability literature, drawing together work by McCarthy (1998), Smith (1999, 2000), Tesar and Smolensky (2000) and Hayes (2004), as well as Prince and Tesar. Most of the arguments are theoretically-driven, but I will also touch on supporting empirical results.

The chapter discusses in detail a proposal for adding a different bias to Biased Constraint Demotion, namely one for the *most specific* IO faithfulness constraint (as in Hayes’ Low-Faithfulness CD algorithm). I discuss the necessity, challenges, and workings of this bias in the second half of the chapter. The prose version of the BCD algorithm that I adopt is given in section 7.1; the reader who understands how that algorithm works need not read anything else here to understand all subsequent chapters.

1.1 Important aspects of Optimality Theory for this learning theory

This dissertation is definitely not the place for the reader to learn Optimality Theory: for introductions to the theory and its primary results, see Prince and Smolensky (1993/2004), Kager (1999) and McCarthy (2002). But I would like to make two introductory remarks about aspects of OT that make it particularly suited to the kind of acquisition questions asked in this dissertation (and that also confront it with the challenges this work tries to address.)

One important aspect of OT for our purposes is the Richness of the Base (ROTB): the strong claim that languages are defined solely on the basis of their constraint rankings, and not by their lexicons, underlying representations or indeed anything else. What this means is that although language users have very different lexicons of stored inputs, every language’s grammar can map *any* possible input onto a legal output – whether a real lexical item, or a nonce word judged grammatical by native speakers. The ROTB assumption places heavy demands on the phonotactic learner: the OT grammar the learner is searching for bears all responsibility for characterizing the languages’ tolerance for marked structures, because none can be shifted onto the lexicon.

Another important learning consequence of the standard OT architecture – in particular the kinds of constraints used and their range of potential conflicts -- is that many rankings can drive the same input-output mapping. This indeterminacy of ranking allowed by each input-output mapping presents the OT phonotactic learner with its main challenge: how to find the ranking that will produce the observed data and nothing more. Such a grammar is termed *restrictive*. As Prince and Tesar (2004) make explicit, the search for restrictiveness is the OT version of the Subset problem (Baker, 1979; Berwick,

1985; Smolensky, 1996), and it is precisely this search that the biases of Biased Constraint Demotion are built to guide.

1.2 Outline of the chapter

Section 2 introduces the main assumptions of OT learning that I adopt, taken most directly from Tesar and Smolensky (2000) and their Constraint Demotion Algorithm and emphasizing the role of stored errors in this learning model. Section 3 introduces Biased Constraint Demotion, and talks about biases that try to attribute marked structures to constraints that will allow for a maximally-restrictive grammar. I demonstrate how Prince and Tesar's BCD algorithm works, using the well-known bias for high-ranking Markedness, and then present the argument from McCarthy (1998) for high-ranking OO-faith (see also Hayes, 2004). Section 4 introduces the trickier issue of how to attribute errors to faithfulness, focusing on the import of specific >> general relations between the contexts of faithfulness constraints, and the problems raised by Prince and Tesar in using these relations to build any simple ranking bias. Leaving aside temporarily these problems, section 5 summarizes a core success of BCD learning: that its reliance on stored errors makes it able to escape superset grammars caused by previous incorrect assumptions.

Section 6 introduces my proposal for imposing the bias for ranking specific faithfulness constraints over general ones, which relies both on constraint definitions as well as the learner's current knowledge of the language. Section 7 synthesizes the results of sections 3 through 6 into the version of BCD that I will use in subsequent chapters; it also provides some more analysis of the ways the specificity of faithfulness constraints

can be calculated and their roles in the search for restrictiveness. Finally, section 8 summarizes the chapter's results, and considers to what extent the BCD model makes empirical predictions about the initial state of acquisition. This discussion will lead us onwards to my novel proposal for a *gradual* BCD learner in chapter 2.

2. Learning an Optimality-Theoretic grammar

2.1 The learning framework: the Tesar/Smolensky learner

The ways in which Optimality Theory differs from previous views of generative (rule-based) phonology also provide OT with its view of language acquisition. To learn a language-specific grammar is simply to learn a constraint ranking; to learn a language means learning that ranking, in tandem with a lexicon of underlying representations.¹

This section provides an overview of the OT learning approach used in this dissertation. For reasons of attribution as well as notational convenience, I will refer to this approach as the Tesar/Smolensky (T/S) view, with reference to their (2000) book – though as we will soon see, my implementation of this view also relies heavily on the learnability contributions of Prince, Hayes, McCarthy and others.

The T/S learner is *error-driven* in all of the following ways: it uses its current grammar to process ambient language data and make errors; it is the making of an error that triggers learning; and it learns by re-ranking constraints in the current grammar, guided in some way by the error.² (As we will see – these errors are just the right ones for

¹ Leaving aside the building of constraints (although see: Hayes, 1999 among others.)

² It is worth pointing out that at least one alternative proposal for learning OT rankings is *error-triggered*, but not really *error-guided*: that of Pulleyblank and Turkel (1998), (2000). Based on previous learnability work in the Principles and Parameters framework (i.e. the Genetic Algorithm approach applied to language acquisition by Clark, 1992), the Pulleyblank and Turkel learner is triggered by the errors it makes, but it selects a new ranking hypothesis through somewhat random re-combinations of rankings rather than any reasoning from the error itself.

our learner to focus on: see also Tesar and Smolensky, 2000, Prince and Tesar, 2004: 257-58.) The rest of this section will illustrate this process.

For the most part, this dissertation is concerned with learning phonotactic distributions – the possible surface structures of the language (segmental inventories and restrictions, syllable shapes, stress patterns, and the like) – and will set aside the further complicating issue of learning phonological alternations.³ So as a first step, we can equip our learner with the Identity Hypothesis⁴ – the hypothesis that the ambient *outputs* that the learner are in fact the *inputs* to the learner’s own grammar.⁵

An error is an optimal candidate under the learner’s current grammar that is not identical to the observed (i.e. heard) winner. Note, however, that the observed winner will in fact match more than one possible output candidate: because although the sound signal contains all the phonetic information about the winner’s features, segments, tones, intonational contours and stress,⁶ the learner must still assign the winner its ‘hidden structure’ – feet, prosodic words, intonational phrases, morphological affiliations and the like. In the present view, learning to assign hidden structure to winners is done in tandem with the acquisition of the rest of the grammar – see the return of this point in sections 4.2-4.3 – but for now we will set it aside.

With the Identity Hypothesis in mind, we can illustrate the T/S learner. Imagine that our learner takes as input a form /A/, provides /A/ to EVAL, and receives as output

³ Although see this chapter §7.3, chapter 3 §6 and chapter 4.

⁴ This assumption runs through the work of Tesar and colleagues; I am not sure whether the term ‘Identity Hypothesis’ originates with Tesar or somewhere else. Hayes (2004) also adopts this assumption, citing a suggestion from Daniel Albrow.

⁵ Another way of stating this hypothesis is simply to invoke Lexicon Optimization (Prince and Smolensky 1993/2004 §9.3): L.O. simply says that the learner will assume an input that is identical to the output, parsed in as unmarked form as possible.

⁶ This dissertation will have nothing to say about how children get from the acoustic signal to all these mental objects – see e.g. Maye (2000); Hayes (1999).

the output [B]. Our current grammar has thus made an error, illustrated in the tableau in 1):

1)

/A/	*A	*C	*B	Ident- A vs. B	Ident- A vs. C
(i) A	*!				
(ii) B			*	*	
(iii) C		*!			*

Our learner’s specific task to establish why it made an error – that is, why its current grammar mapped /A/ to [B], and not to [A] – so we can ignore the rest of the candidate set and just compare the two output candidates [A] and [B]. In the version of the T/S learner that I adopt, this comparison is represented in a distilled form as in 2):

2) Boiling down the information in tableau 1)

/A/	*A	*B	*C	Ident- A vs. B	Ident- A vs. C
A ~ B	L	W	e	W	e

In Prince (2002), this distillation of candidate comparisons is called an Elementary Ranking Condition vector (Prince 2002a,b.) In this dissertation, I will refer to objects like 2) as ERC rows. What each cell in an ERC row tells us is the preference of each constraint with respect to the winner and its rival loser candidate. In this case: the tableau in 1) shows that *A assigns a violation mark to the winner [A], and no mark to the loser [B], so we can say that **A prefers the loser*: thus the ERC row for the A~B comparison contains an L in the *A column. Since the second markedness constraint in the table *B assigns the opposite violation marks (one to [B] and none to [A]), **B prefers*

the winner: this puts a **W** in its column. The third markedness constraint *C assigns equal violation marks (in this case, none) to both the winner and loser candidates: thus, *it prefers both winner and loser equally*, and this equality puts an **e** in the *C column.

In this way the Ls, Ws and es of an ERC row indicate the relevant discrepancies between the current and target grammars. The core of T/S learning is to reason from these discrepancies to necessary rankings, and the logic of OT gives us the following way to characterize them. This logic is given as the Cancellation/Domination Lemma of Prince and Smolensky 1993: 148; rephrased like this in Prince and Tesar 2004: 255:

- 3) If *every* L-prefering constraint is ranked below *some* W-prefering constraint, our grammar will prefer the Winner to the Loser.

This lemma is the crux of the recursive Constraint Demotion Algorithm (CDA: Tesar, 1995; Tesar and Smolensky, 1998, 2000; see also Prince 2002a,b). The CDA is a technique for modifying a ranking, by demoting Loser-prefering constraints until the ranking no longer makes the errors encoded in its current ERC rows. We will soon see how this works below.

This logic also drives the class of Biased Constraint Demotion algorithms (Prince and Tesar, 2004), which I will adopt in this dissertation. In BCD, each cycle of learning (i.e. constraint re-ranking) creates a new grammar hypothesis, and this new grammar will cause a new set of errors and consequent mark-data pairs. While previous grammars are forgotten as soon as a new one is built, the T/S learner I will use *retains* its ERC rows, in a table called the Support. The re-ranking algorithm always works with reference to the Support: the sum of all errors the learner has ever made.

2.2 The importance of errors and the Support

Tesar and Smolensky (2000) provide a proof that that recursive application of the CDA will take any Support and successfully find a ranking that ‘resolves’ all the Support’s errors (assuming one exists.) A ranking that ‘resolves’ a set of errors is one that chooses all the winners instead of their respective losers. In this system, the first fundamental role of the Support is to provide the algorithm with the errors to learn from.

The Support is also crucial to tackling many subparts of the learning problem with a BCD learner. Tesar (1997, 1998) points out that a memory for errors like the Support allows the learner to do what he calls Inconsistency Detection, which means noticing that no one ranking exists to describe the data. Detecting inconsistency is a specialty of the Support, and it has many functions: see the applications of Inconsistency Detection in e.g. Prince (2002); Tesar et al (2003); McCarthy (2005).

One application is the matter of learning “hidden structure”: properties of winners that the sound signal does not carry, and which the learner must therefore infer. Such structure includes both prosodic and morphological information – syllabification, footing and higher-level prosodic structure, as well as morphological category and paradigmhood – and also underlying representations, which may turn out to not match the observed outputs.

As another example: the Support provides an approach to learning an OT grammar that is sensitive to exceptions and/or lexical strata. This is the case in the hypothetical Support given below, taken from Pater (to appear), in which *every constraint prefers some loser*:

4) *A Support that is inconsistent*

winner ~ loser	NoCoda	Max
pa ~ pak	W	L
lo ~ lok	W	L
tak ~ ta	L	W

Pater (to appear) uses Inconsistency Detection to learn a grammar that encodes exceptionality through lexically-indexed constraints. When faced with a Support like 4), this learner finds a constraint that prefers no losers *for all instances of some morpheme*, and then installs a version of that constraint indexed to all the morphemes for which it favours only winners. (See also the discussion of exceptionality in Winslow (2003) and Pater (2004b). I return to the case of exceptionality in chapter 3 §6.

The Support is also very crucial to the proposals made in this dissertation. Later in this chapter (§6), I propose how the Support should be used to calculate contingent ranking biases. In the next chapter, I propose a novel way by which errors get into the Support, which derives intermediate stages of acquisition (chapter 2 §3), and also some of the variation between those stages (chapter 2 §6.)

2.2.1 The Support and the lexicon

One question that arises in the attempt to use the T/S learner in real-life learning is the connection between the Support and the phonological lexicon. In some ways, it may seem that the Support should in fact be considered as a proto-lexicon, since it contains observed words of the language that children must be learning, and since as we will see the learner must update their entries in the Support as they learn more about e.g. the morphological structure of those words.

Nevertheless, the Support as it stands contains both more and less information than a phonological lexicon. On the one hand, the Support contains not just the language's observed outputs but also their associated losers and comparisons of violation profiles. On the other hand, the Support only contains those forms that induced errors so it will not include lexical items that are faithfully parseable under the current grammar. Furthermore, the Error-Selective proposal that I make in the next chapter is very attuned to the purely phonological properties of the errors that are added to the Support, and in what order, with absolutely no concern for whether a child has learned the meaning or lexical quirks of any particular error-inducing word. Thus, it seems that the Support and the lexicon are two different mental objects, entrusted with the storage of different knowledge; the relationship between them is an unresolved issue.

2.2.2 The Support and the Gradual Learning Algorithm

The most well-known alternative approach to learning OT is the Gradual Learning Algorithm (e.g. Boersma, 1997; Boersma and Hayes, 2001; Boersma and Levelt, 2000; Curtin and Zuraw, 2001; Levelt, and van der Vijver, 2004.) Although there are other key differences between GLA learning and what I've discussed above – a basic assumption of the GLA is that it has no analogous notion to the Support. Learning re-ranks constraints gradually on the basis of one error at a time, and errors are never stored. This means that the GLA is not equipped to handle the learning problems of hidden structure addressed above in a consistent way.⁷ This difficulty with the GLA is the focus of chapter 3§5-6.

⁷ See Boersma and Appousidou (2003, 2004), where the GLA succeeds in learning metrical structure on some trials but not others.

3. Restrictive phonotactic learning: Biased Constraint Demotion

Since the T/S learner is error-driven, it continues re-ranking constraints until it stops making errors. So, the crucial test of the T/S learner is that when it stops making errors, its constraint ranking is indeed the target ranking – that is, that the ranking embodies all the properties that the analyst ascribes to native adult speakers. (There is also the considerable issue of whether the learner’s inputs are also correct; which I ignore for the present although will touch briefly on in chapter 4.)

As discussed in section 1, many different constraint rankings will choose the same optimal input for a given output – which also means that each ERC row will only partially determine the nature of the new grammar to be learned. In choosing between these rankings, the OT learner runs into the classic learnability problem of subset and superset grammars, which I present below.

Given the re-ranking logic given in 3) above, the single ERC row repeated below in 5) will be resolved by any grammar that includes *at least one* of the rankings in 6):

5) *One ERC*

<i>input</i>	<i>winner ~ loser</i>	*A	*C	*B	Ident-A/B	Ident-A/C
/A/	A ~ B	L	e	W	W	e

- 6) *The rankings that resolve this ERC*
 (a) *B >> *A OR
 (b) Ident-A/B >> *A

Clearly, many *total* rankings of these 5 constraints will contain one or the other of the partial rankings in 6), e.g.:

- 7) (a) *C >> *B >> *A >> Ident-A/B >> Ident-A/C
 (b) Ident-A/B >> Ident-A/C >> *A >> *B >> *C
 (c) *C >> Ident A/C >> *B >> Ident-A/B >> *A

The crucial concern in choosing between these rankings is that our learner must not only choose a grammar to map all the /A/s to themselves instead of to [B]s. There is also the need to rule out any other unattested surface forms, [C]s through [Z]s – and as we will see shortly, some of our the rankings in (7) do this job better than others.

To use a concrete example, consider the stress rule of French, in which main stress falls on the last syllable of the word. Imagine a French learner makes the error in 8a), and so creates the ERC row in 8b):

8)a)

/papá/	Trochee	Iamb	Non Finality	Ident-Stress
(i) (pa.pá)	*		*!	
(ii) (pá.pa)		*		*

8)b)

<i>input</i>	<i>winner ~ loser</i>	Trochee	Iamb	Non Finality	Ident-Stress
/papá/	(papá) ~ (pápa)	L	W	L	W

If our learner decides to resolve this error by installing Ident-Stress above the two L-preferring constraints, the resulting grammar will be as in 9) below:

9)

/papá/	Ident-Stress	Trochee	Iamb	Non Finality
(i) (pa.pá)		*		*
(ii) (pá.pa)	*!		*	

By ranking Ident-Faith highest, the learner has indeed ensured that the final stress of any input French word is preserved in the output. But it has also learned a grammar in which stress is *lexically* determined – merely falling wherever it is in the input. And the generative assumption is that this is NOT the grammar that French speakers have learned – it is not just a fact (or accident) of the lexicon that every single French word has ultimate stress, but also a fact we want to attribute to (and capture in) the phonological grammar of the French speaker.

From the OT perspective, we can say that our learner is searching for a grammar that faithfully reproduces all the attested forms, and also maps all of the rich base onto attested forms. In other words, we want the learner to find the most *restrictive* grammar.

This issue is clearly not new to OT. Much linguistic learnability work has centered on ensuring restrictiveness, and it has driven various proposals about the nature of the grammar itself.⁸ As we saw in the French stress example above, the error-driven OT learner suffers from the well-known subset problem because of a lack of positive evidence. The danger of relying on the current grammar to provide errors to learn from is that learners will never make errors that show they've chosen an insufficiently restrictive grammar. So the two goals are first to identify what makes the most restrictive grammar (constraint ranking) among any set of options consistent with the data, and then to ensure that the learner learns that ranking.

As already cited, the search for restrictiveness in OT learning is at the core of the proposals in Prince and Tesar (2004) and also Hayes (2004) – in what follows, I start from the Prince and Tesar model of BCD, but I adopt insights and technology from both works. The central idea of BCD is to give the learner a set of prior assumptions about

⁸ See e.g. Dresher and Kaye (1990); Dresher (1999).

constraint rankings, called ranking biases, which the learner assumes up until the learning data provides evidence to the contrary. Building a constraint ranking is a series of cycles of adding constraints to strata – starting at the top and continuing until there are no more constraints to be ranked. In building each stratum, the learner aims to install all constraints that its biases want highest-ranked, and put off the installation of all other constraints until it has to.

To see how Biased Constraint Demotion works, I will start with an illustration of the most basic for choosing the most restrictive OT grammar: M >> F.

3.1 Illustrating the BCD approach: high-ranking Markedness

In the terms of OT learnability, the Markedness >> Faithfulness bias first appears in Smolensky (1996) and Tesar and Smolensky (1998) (elaborating on a suggestion made by Alan Prince.) In the literature on children's productions, this observation goes back at least to Jakobson (1941/1968); see also the works of e.g. Jakobson and Halle (1956), Stampe (1969), Macken (1978), Dinnsen (1992); Fikkert (1994); and in the OT context, Gnanadesikan (1995), Demuth (1995), Pater (1997) *inter alios*. The more Markedness is high and Faithfulness is low in a grammar, the fewer marked surface structures it permits in the language, and thus the more restrictive the grammar is. So if the constraint that the learner is going to use to choose a loser over a winner could either be an M or an F constraint, the drive for restrictiveness should make them choose the M constraint.

Let us see how the BCD approach enforces M >> F. Note that while we are only using one bias here, the same reasoning will apply no matter how many biases we add

into the model. In this example, we will assume our learner has only added one error to the Support, being the schematic error we've already seen:

10) *the Support – a collection of ERCs*

input	winner ~ loser	*A	*C	*B	Ident-A/B	Ident-A/C
/A/	A ~ B	L	e	W	W	e

What we've already seen is that any ranking with either *B or Ident-A/B >> *A will get the winner to be more harmonic than the loser; what we know now is that we want to choose the *B >> *A ranking. And we now have two competing goals: resolving the error in 10) and respecting the M >> F ranking bias.

Remembering the logic of CDA: resolving errors means installing *some* W-preferring constraint over *every* L-preferring constraint. Once this has been done for any particular ERC row, its loser is guaranteed to be less optimal than its winner, and so that error can be ignored for the rest of the ranking process. So the BCD imposes the M >> F bias by first installing all *M constraints that do not prefer the loser*, and then checking whether the error has been resolved:

11) *Step 1:* Install all M constraints that prefer no losers
Resulting stratum 1: *B, *C

Looking back at our MDP in 10), we can see that our error has indeed been resolved, because one of the installed constraints, *B, prefers the winner, so:

12) Remove from consideration all resolved errors
Resulting Support: -- empty --

Now we go through the second cycle, to add constraints into stratum 2. Since there are no remaining errors, there are no constraints to prefer any losers, so our bias is free to install all the remaining markedness constraints:

13) *Step 1:* Install all M constraints that prefer no losers
Resulting stratum 2: *A

14) Full ranking so far: *B, *C >> *A

(This also means we have no errors to remove in part 2.)

Now we have installed all the constraints that our bias wants to rank high. This means we are safe to dump all the remaining constraints (the faithfulness constraints) at the bottom of the hierarchy:

15) *Step 1:* Install all M constraints that prefer no losers
Resulting stratum 3: -- empty --

Step 2: Install all remaining constraints in the last stratum (*to be revised*)⁹
Resulting stratum 3: Ident A/B, Ident A/C

16) Final full ranking: *B, *C >> *A >> Ident A/B, Ident A/C

Happily, BCD has found the right ranking: the ranking in 15) chooses the winner [A] over its rival [B]:

⁹ The discussion of how faithfulness constraints should be installed, even when errors still remain, will be extensive – see §4, 6-7 below.

17)a) *The correctness of *B >> *A*

/A/	*B	*C	*A	Ident A/B	Ident A/C
(i) A			*		
(ii) B	*!			*	

And because we chose the *B >> *A ranking to get our winner to beat its rival loser, the learner has not made any unmotivated concessions to faithfulness. So, for example, if we were now to encounter an input that has features protected by Ident A/B, it will still be neutralized to something less marked.

17)b) *The restrictiveness of *B >> *A*

/B/ ¹⁰	*B	*C	*A	Ident A/B	Ident A/C
(i) A			*	*	
(ii) B	*!				

A note: it will already be clear that BCD does not build completely classic OT grammars in one sense, because the ranking in 15) is a stratified, partial ordering of constraints rather than a total ordering – but any total constraint ordering that is consistent with this partial ordering can be selected.

Stepping back: we can say that the role of the M >> F bias is to put as little burden on the lexicon and as much on the grammar as possible. In other words, it chooses rankings that will rule out as many unseen forms as possible, without requiring negative evidence of their non-existence. Of course, it will soon come to pass that a learner’s positive evidence cannot be resolved by installing Markedness constraints alone. The

¹⁰ This input is provided only to demonstrate the restrictiveness of our ranking in winnowing down the Rich Base. If this input were actually the product of real phonotactic learning – that is, the learner was hearing outputs [B]s – this ranking would be wrong, but the point here is what our learner has decided only on the basis of output [A]s.

second ranking bias which I integrate into Biased Constraint Demotion provides some additional help: this is the bias for high-ranking OO-Faith.

3.2 A second bias in BCD terms: high-ranking OO-Faith

Although the central notion of faithfulness of Prince and Smolensky (1993/2004) is between input and output (IO-faith), work in OT since then has adopted a variety of different faithfulness relations – e.g. between bases and reduplicants (McCarthy and Prince, 1995 *et seq.*), and also between morphologically-simple and derived outputs. Here I will focus on this latter relation.

As the name suggests, OO-faith constraints assess similarity *among* output forms. Output-output relations are the OT answer to the long-observed phenomenon of paradigm uniformity: i.e., that the phonological regularities of a language are often overridden just where they would cause morphologically-derived forms to differ from their bases. In other words: some phonological generalizations only have exceptions that keep the derived forms of a morphological paradigm similar to their base. In the spirit of this wording, the OT accounts of paradigm uniformity (defined variously: see Burzio 1997, 2000; Kenstowicz, 1997; Kager, 1998; Steriade 1998, 2000 *inter alia*) all enforce something akin to faithfulness between morphologically-related surface forms. The choice of proposals is not crucial -- as far as I know, the learnability arguments to follow do not hinge on any particular account of paradigm uniformity.

3.2.1 OO faith as an OT account of cyclicity

One famous example of the phenomenon is the interaction of flapping and Canadian raising (CR) in some dialects of English (e.g. Joos 1942; Chambers 1973; Mielke et al 2003).¹¹ In such dialects, CR is purely allophonic in monomorphemic words: raised [ɔɪ] appears before voiceless obstruents as in ‘write’ [ɹaɪt], while [aɪ] appears elsewhere as in ‘ride’ [ɹaɪd]. However, derived forms with a base vowel [ɔɪ] exceptionally retain their raised quality even before a voiced flap, as in ‘writer’ [ɹaɪrɪ], * [ɹaɪrɪ]. In other words: diphthongs are unraised before flaps *except when they are raised in the base*.¹²

- 18) *Exceptional Canadian Raising in words derived from raised bases*
 rider, [ɹaɪrɪ] vs. writer, [ɹaɪrɪ] (c.f. ‘write’, [ɹaɪt])
 wider, [waɪrɪ] whiter, [waɪrɪ] (c.f. ‘white’ [waɪt])

The constraint set I will use here: *Output-Output Faithfulness* (Benua 1997, 2000)

- 19) Output-Output-Faith-[F], informally defined
 “Derived words must match their base’s value for the feature [F]”

The schematic analysis of such a pattern will be OO-Faith >> Mark >> IO-Faith. First, Mark >> IO-Faith ensures the normal distribution on raised diphthongs; in this environment (i.e. before a voiced flap) the context-free markedness constraint against

¹¹ By ‘some dialects’, I refer to one of the two dialects originally reported in Joos (1942). Whether this particular dialect is anything more than an idealization among modern-day speakers is a separate question: see especially Hall (2005).

¹² In fact, it appears that the intervocalic flapping process that creates the environment for exception raising in ‘writer’ is in fact *also* OO-Faith sensitive in longer words -- see Wittgott, 1982; Steriade, 2000; Davis 2002 on ‘capi[r]alistic’ vs. ‘mili[t]aristic’.

raised diphthongs rules [ɔɪ] out (see 21a below). OO-faith >> M enforces exceptional raising: an undominated OO-faithfulness constraint that regulates vowel height (OO-Ident-[hi]) will protect raised diphthongs only to keep derived forms similar to their bases (21)b):

- 20) *The constraints*
- | | |
|---------------|--|
| *ɔɪ | No raised diphthongs |
| OO-Ident-[hi] | Vowels in <i>derived</i> outputs must match their output <i>bases</i> correspondent’s value for the feature [hi] |
| IO-Ident-[hi] | Vowels in outputs must match their input correspondent’s value for the feature [hi] |

21) The rankings

(a)

‘rider’	*ɔɪ	IO-Ident [hi]
(i) ɹaɪrɪ	*!	
(ii) ɹaɪrɪ		*

(b)

/ɹaɪt + ɹ/ ¹³	OO-Ident [hi]	*ɔɪ	IO-Ident [hi]
(i) ɹaɪrɪ		*	
(ii) ɹaɪrɪ	*!		*

Examples of this phenomena are robustly attested across languages – famously, stress in derived words is often constrained by paradigm uniformity (on Arabic, see e.g. Brame, 1974, Kager, 1999; on English stress, see e.g. Chomsky and Halle, 1968; Pater, 2000.) Two other, different examples are illustrated below:

¹³ In this tableau, I have given the output form of the base, ɹaɪt, in the underlying form, to show what OO-Ident is being faithful to. However, I have not made explicit how the OO-faithfulness constraint knows what the base’s surface form is – here I am just assuming that this is possible. In Benua’s (1997, 2000) model, this is done by actually running the base through EVAL as part of the evaluation of the derived form. As mentioned in the text: the way we implement OT paradigm uniformity will affect the form of the base and the tableaux themselves – but not, I think, the learnability result.

- 22) English sonorants are syllabified as onsets before vowels *except when they are syllabified as nuclei in the base*:

light.ning	vs.	<u>ligh</u> .ten.ing	(c.f. <u>ligh</u> .ten)
bright.ness		<u>brigh</u> .ten.ing	(c.f. <u>brigh</u> .ten)
William Faulk.ner		<u>fal</u> .con.er	(c.f. <u>fal</u> .con)
Hugh Heff.ner		<u>oft</u> .en.er	(c.f. <u>oft</u> .en)

- 23) In Sundanese, nasalization does not spread across an oral consonant *except when the target of spreading is nasalized in the base* (Robins, 1957; Cohn, 1990; Walker, 1998)

(a) *f*+nasal] spreads rightwards only through vowels/glottals

[n̄īar]	‘seek’
[m̄ah̄al]	‘expensive’
[bīh̄ar]	‘to be rich’

(b) *spreading blocked by [r] and [l]*

[ŋ̄ūliat]	‘stretch’
[m̄arios]	‘examine’

(c) *but base vowels still nasalized even across the infix [-a] / [-ar]*

[ŋ̄-āl-īar]	‘seek, plural’
[m̄-ār-āh̄āl]	‘expensive, plural’

3.2.2 OO-faith as an OT account of MSCs

A different use of OO-faith, relevant to the learning discussion to follow, is McCarthy (1998)’s OT reanalysis of Morpheme Structure Constraints (see e.g. Chomsky and Halle 1968; Kisseberth 1970.) McCarthy uses the example of the distribution of root vowel length and Minimal Words in the ‘Kansai B’ dialect of Japanese. In this dialect, there is a static generalization that roots are always at least bimoraic (e.g. [kaa], *[ka]) – this is true independent of any surrounding morphology.

- 24) Kansai /B/

<i>possible paradigms:</i>		<i>impossible paradigms:</i>	
<i>root</i>	<i>root + affix</i>	<i>root</i>	<i>root + affix</i>
(a) [kaa]	[kaaga]	(b) *[ka]	[kaga]
		(c) *[kaa]	[kaga]

Assuming that the bimoraicity minimum is the work of the high-ranking markedness constraint FootBinarity (Prince and Smolensky, 1993), we can explain why paradigms like (24b) do not occur, because FtBin >> IO faithfulness to syllable weight will force any input mono-moraic root to lengthen on the surface:

- 25)

/ka/	FtBin	IO-Ident-Wt
(i) ka	*!	
(ii) [∞] kaa		*

But if we can have input roots like /ka/, why are there no surface paradigms like (24c)? That is: what forces a root to *remain* lengthened, even in derived forms where word minimality is no longer at issue? McCarthy’s point is that paradigms that alternate can be ruled out with OO-faith to syllable weight (i.e. moras). A grammar in which OO-Ident-Wt outranks its IO-faith counterpart will give us this result:

- 26)

/ka + ga/ base: [kaa]	Ft-Bin	OO-Ident-Wt	IO-Ident-Wt
(i) kaga		*!	
(ii) [∞] kaaga			*

The role of OO-faith here is to ensure that derived forms match their bases for vowel length -- even if Markedness does not require long vowels in that environment.

3.2.3 The learnability argument for high-ranking OO-faith: McCarthy (1998)

The learnability argument for high-ranking OO-Faith bias also comes from McCarthy's discussion above. According to the ranking in (26), Markedness ensures that roots are bimoraic in simple words, while OO-Faith insists that the root portion of a complex word remain similarly bimoraic. How can the learner get to this ranking? While the M >> F bias means that the initial state already contains the ranking that lengthens hypothetical roots in (25), nothing in the data will drive the ranking between OO- and IO-Faith necessary in (26).

To restate the argument in the terms of our BCD model: this learner of the non-alternating dialect will only have evidence for inputs like /kaa/ and /kaaga/, and so only have the two ERCs in (27) (note that I have included the M constraint *LongV in this Support table, to explain why the learner might make these errors in the first place):

27) *ERCs for McCarthy's Kansai B example*

input	winner ~ loser	FtBin	*Long-V	OO-Ident-Wt	IO-Ident-Wt
/kaa/	kaa ~ ka	W	L	e	W
/kaa + ga/	kaaga ~ kaga	e	L	W	W

This Support merely tells the BCD learner that *LongV needs to be demoted below some winner-preferring constraints – and that given the error in the derived context, *LongV must get below *either* OO or IO faith. McCarthy's point is thus that a ranking bias for OO >> IO faith is necessary. This bias will ensure that OO-Ident[hi] is installed high and therefore exclude the possibility of alternating paradigms like (24c).

This example provides the restrictiveness argument for ranking OO-faith over IO-faith; it says nothing about its interaction with Markedness. In the version of BCD I will

put together at the end of this chapter (§7.1), however, OO-faith will in fact be installed above Markedness constraints wherever possible as well. The arguments I will provide for this choice will come from empirical evidence from developmental stages in the literature, noted by Hayes (2004) and others – this data will be discussed in chapter 4 7.3.1. However, one can also demonstrate that an OO-Faith >> Markedness bias is required to ensure the acquisition of some end-state grammars as well: see Becker (2006).

3.3 Connecting M >> F and OO >> IO as surface-oriented biases

Bruce Tesar (p.c.) points out that the high-ranking M and OO-F biases have in common a reliance on surface evidence. In other words: the violation profiles of M and OO-faith are defined exclusively by looking at the outputs that children have not just hypothesized, but actually heard. As a result, their full potential for responsibility for surface contrasts and neutralizations is already known.¹⁴

One related point is that this surface-oriented property makes it straightforward to learn the appropriate *context* of markedness and OO-faith activity using error-driven learning. To see this, I show here how our M >> F biased learner correctly acquires a grammars in which a specific and general markedness constraint are crucially ranked with respect to one another.

This very simple example is the case of a language in which coda consonants are allowed, but coda *clusters* are not permitted (Blevins 1995 lists the languages Thargari, Sedang and Mokilese as having this syllable profile.) The analysis of this pattern that I will adopt in chapter 2 is one with two markedness constraints in a stringency relation: a

¹⁴ The one hypothesis that the learner must make in the case of OO-faith is to determine which segments in a derived form make up the morphological base.

general NoCoda constraint, and a more specific NoComplexCoda constraint. The singleton coda pattern comes from sandwiching faithfulness between these markedness constraints. If we assume for example that coda clusters would be repaired via deletion, we can use the ranking in 28a) below to derive the right results:

- 28) *The singleton coda grammar*
 (a) NoComplexCoda >> Max >> NoCoda

(b) Faith >> General M			(c) Specific-M >> Faith			
/bab/	Max	NoCoda	/balb/	*Complex Coda	Max	NoCoda
(i) bab		*	(i) balb	*!		*
(ii) ba	*!		(ii) bab		*	

The error-driven BCD learner will discover the correct ranking in 28a) without incident. Our phonotactic learner will only make one kind of error – one which reduces the singleton codas it hears in words like [bab] – and this error will produce the ERC in 29) below:

- 29) *ERCs for the singleton coda grammar*

winner ~ loser	*ComplexCoda	NoCoda	Max
bab ~ ba	e	L	W

On the first pass of learning from this error, the high-ranking M bias from section 3.1 will first install the specific *ComplexCoda in the top stratum, simply because it does not prefer the loser. Since it does not prefer the winner either, though, it will not resolve the error, so BCD will be forced to install faithfulness in the second stratum and then place NoCoda at the bottom of the ranking. This process generates the correct ranking from 29a).

What is perhaps less obvious is that this ease of learning holds of OO-faith as well. That is, if various OO-faith constraints with different contexts can explain the surface structure of derived forms, the learner's errors will demonstrate which ones should be demoted. The following illustration comes from the famous example of cyclic effects in Palestinian Arabic (Brame 1974), in which just those vowels with main stress in a morphological base escape vowel syncope in derived forms: the data are summarized in 30) below.

- 30) a) *In Palestinian Arabic, unstressed [i] is usually deleted:*

/fihim-u/_{subj} → [fihmu], *[fihimu]
 /fihim-na/_{subj} → [fihimna], *[fihimna]

- b) *even if the [i] comes from a morphological base (bases underlined):*

/fihim/ → [fihim] and: /fihim-u/_{acc} → [fihmu], *[fihimu]

- c) *... except when the base [i] was stressed:*

/fihim/ → [fihim] so: /fihim-na/_{acc} → [fihimna], *[fihimna]

This effect is re-analyzed by Kager (1999)¹⁵ using OO-Max constraints relativized to segments in the base's prosodic head (see also McCarthy, 1995a; Alderete, 1999). The relevant constraint, in both Kager's definition and the current OO-faith terms, is in 31):

- 31) Head-Max(B/O): 'Every segment in the base *prosodic head* has a correspondent in the output' (Kager, 1999: 214)
in other words:
 OO-Max[Seg]-prosodic head

¹⁵ See also Kenstowicz, 1997; Steriade, 1998.

In comparing the base and derived forms in 30b) and c), the learner gets the overt evidence that OO-Max-Prosodic Head is satisfied while general OO-Max is violated (as well as IO-Max):

32) *An OO-faithful grammar*

(a) $M \gg$ General-OO-faith

$\begin{matrix} /fihim-u/_{acc}, \\ [fihim]_{base} \end{matrix}$	Syncope	OO-Max[Seg]
(i) φ <u>fihmu</u>		*
(ii) <u>fihimu</u>	*!	

(b) Specific-OO-faith \gg $M \gg$ General-OO-faith

$\begin{matrix} /fihim-na/_{acc}, \\ [fihim]_{base} \end{matrix}$	OO-Max[Seg] Prosodic Head	Syncope	OO-Max[Seg]
(i) φ <u>fihimna</u>		*	
(ii) <u>fihimna</u>	*!		*

With these violation profiles, the learner does not need any further bias apart from its preference for OO-faith constraints, and for constraints that prefer no losers, to get the right ranking. Since we are assuming an initial grammar where OO-faith is undominated, the learner will only be making errors like 33a):

33) *ERC row for the Palestinian Arabic syncope grammar*

<i>input</i>	<i>winner ~ loser</i>	OO-Max[Seg] Prosodic Head	Syncope	OO-Max[Seg]
$\begin{matrix} /fihim-u/_{acc}, \\ [fihim]_{base} \end{matrix}$	<u>fihmu</u> ~ <u>fihimu</u>	e	W	L

This error makes it clear that Syncope must rank above the general OO-Max constraint; this will produce the ranking 32b).

3.4 **Interim Summary**

This section has demonstrated the idea of Biased Constraint Demotion, with two biases for ranking surface-oriented constraints (Markedness and OO-faith) above IO-faith. The resulting ranking biases our learner uses install as many of these constraints as possible at every stratum, resolving as many errors as possible without resorting to assumptions about the input.

With this backbone of BCD in place, we're left with the much less straightforward issue of the ranking choices between IO-faithfulness constraints. When no further errors can be resolved without installing some IO-faith constraints, how should the learner choose among them? Which choices will consistently result in a more restrictive grammar?

4. **An input-oriented ranking bias in BCD: Specific-F \gg General-F**

Prince and Tesar (2004) argue compellingly for principles that choose between installable faithfulness constraints during phonotactic learning with an eye to ensuring restrictiveness. While I do not argue with the usefulness of their principles in the cases they discuss, what I do here is concentrate on one bias they do *not* adopt, namely a bias for installing the most specific faithfulness constraint (Smith, 1999, 2000; Hayes, 2004). To understand this bias, this section introduces a set of specific faithfulness constraints – positional faithfulness – and make the case for both this bias' necessity and its challenges.

4.1 The theory of positional faithfulness

The positional view of faithfulness constraints that I adopt takes as its starting point a general observation that is so well summed up in Smith (2002) that I quote verbatim:

- 34) “There is a set of phonologically prominent or “strong” positions that are well known for their special ability to license phonological contrasts, resisting neutralization processes that may otherwise be active in a language (Trubetzkoy 1939; Steriade 1993, 1995, 1997; Beckman 1995, 1997, 1998; Casali, 1997; Padgett 1995; Lombardi 1999; Zoll 1996, 1997, 1998) [...] Many languages will tolerate a particular phonological contrast, such as that between voiced and voiceless obstruents or that between oral and nasal vowels, only inside one of these strong positions. Specific examples of special contrast-licensing behavior in the various strong positions can be found in the references cited above. “(Smith 2002: 8).

Further references with relevant data and arguments include Kingston, 1985; Lombardi, 1991, 1996; Selkirk, 1994; Alderete, 1995, 1999; Smith, 2001, 2002.

Positional faithfulness constraints are thus an encapsulation of this “special ability to license phonological contrasts”: they formalize the claim that it’s not an accident that some languages contrast voicing in onset and neutralize it in coda, but not the other way around. A fairly comprehensive set of such positions are used to define constraints in 35) below – here I use them to create a series of Ident[voice] constraints:

35) *A set of Ident[voice] constraints*

- a) Ident[vce]-segment “Output segments must match their input correspondents must match for the feature [vce]”
- b) Ident[vce]-V: “Output *long vowel* segments must match their input correspondents for the feature [vce]”

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

- c) Ident[vce]-Onset¹⁶ “Output segments *in syllable onsets* must match their input correspondents for the feature [vce]”
- d) Ident[vce]-σ “Output segments *in initial syllables* must match their input correspondents for the feature [vce]”
- e) Ident[vce]-⁻σ “Output segments *in stressed syllables* must match their input correspondents for the feature [vce]”
- f) Ident[vce]-Root “Output segments *in morphological roots* must match their input correspondents for the feature [vce]”
- g) Ident[vce]-Noun “Output segments *in nouns* must match their input correspondents for the feature [vce]”

These Ident constraints have been defined with respect to *output* contexts. As we will see in chapter 2, other contextual faithfulness constraints – at least positional Max -- must be differently defined, either by referring to *input* contexts or through some other mechanism. These definitional issues are not the focus of this work, but I will point out the definitional assumptions necessary to my analyses when they arise.

4.1.1 Why not (only) positional markedness

This dissertation does not argue that positional faithfulness constraints are the *only* way the grammar encodes contextual sensitivities – that is, that there are no positional markedness constraints. However, neither do I adopt the position of Prince and Tesar (2004) that positional *markedness* constraints should be the only way, in virtue of their ease in learning. Instead, I claim that some positional faithfulness constraints are

¹⁶ It has been suggested that the proper context of this constraint is ‘released consonants’ or something similarly phonetic in its definition (see e.g. Kingston, 1985; Steriade, 1999; Côté, 2000.) In this work, however, I will continue to use the syllabic position Onset.

indeed necessary to capture the range of both developing and adult grammars, and therefore that their learning consequences must be taken seriously.

To support this claim, this section puts the learnability arguments of this chapter on hold and provides three arguments in favour of positional faithfulness.¹⁷ These arguments are (i) its ability to capture the similarity between positional neutralization and assimilation, (ii) its ability to characterize strong positions as blockers, as pointed out in Beckman (1998) and (iii) its ability to capture generalizations about positions whose complements seem to be non-categories.

The first argument comes to me from Pater (p.c.); it originates in part in Mester and Ito (1989)'s analysis of onset-driven voicing assimilation, and was taken up in the OT literature by Cho (1990) and Lombardi (1991, 1996, 1999). The relevant generalization is that onsets both preferentially resist obstruent voicing neutralization as compared to codas, and also preferentially determine the value of coda-onset voicing assimilation. As cited to this end by Lombardi (1996), languages like Polish, Dutch, Catalan and Sanskrit demonstrate this privilege of onset voicing, in that the voice specification of their obstruent clusters is determined by the input voicing of the onset segment and their word-final segments are uniformly voiceless. As 36) illustrates, the positional faithfulness constraint Ident-Onset provides a unified account of both cross-linguistic tendencies:

¹⁷ The reader who does not wish to lose the thread of the learning argument and is willing to grant the existence of such constraints is advised to skip to §4.1.2 below.

36) *The positional faithfulness account*

a) *coda (word-final) neutralization*

/bad/	Id[vce] -Onset	*VcdObs	Id[vce]
(i) bad		**!	
(ii) φ bat		*	*
(iii) pat	*!		**

b) *onset-driven cluster assimilation...*

/adpa/	Agree [vce]	Id[vce] -Onset	Id [vce]
(i) adpa	*!		
(ii) φ atpa			*
(iii) adba		*!	*

(c) *... to either voicing value*

/atba/	Agree [vce]	Id[vce] -Ons	Id [vce]
(i) atba	*!		
(ii) φ adba			*
(iii) atpa		*!	*

On the other hand, a positional markedness constraint like *Coda-VoicedObstruent does not provide the same connection. While this constraint can also explain contextual neutralization as in 37a) below, it cannot explain why a language would ever resolve a coda-onset mismatch by becoming uniformly *voiced*:

37) *The positional markedness account:*

a) *coda (word-final) neutralization*

/bad/	*Coda-VcdObs	Id[vce]	*VcdObs
(i) bad	*!		
(ii) φ bat		*	*
(iii) pat		**!	

b) *... but not coda-to-onset assimilation (cf. 36c with winner [adba])*

/atba/	Agree[vce]	Id[vce]	*Coda-VcdObs
(i) atba	*!		
(ii) \ominus adba		*	*!
(iii) φ atpa		*	

This argument also demonstrates the advantage of faithfulness constraints relativized to *featural* contexts as well. Similar to the onset/coda case above, it is also the case that stops both resist place neutralization over nasals, and also determine the direction of nasal/stop place assimilation (Joe Pater, p.c.). See also especially Steriade (2000) on the special behaviour of retroflex consonants in both direction of assimilation and positional neutralization compared to other places of articulation. I will return to featurally-limited faithfulness of this sort in §6.1.

The second argument is the ability of segments in privileged contexts to block phonological processes, in a way that cannot be expressed in terms of markedness. The compelling example of the blocking comes from Beckman's discussion of Guaraní vowel harmony (Beckman 1998: 153-184.) The relevant distribution of Guaraní nasality can be described as follows: (a) stressed vowels freely contrast for nasality (they are either nasal or oral), and (b) unstressed vowels and sonorants are only nasalized through a process of nasal assimilation, in which nasality spreads leftward from a stressed nasal vowel, nasalizing all sonorants and passing transparently through voiceless obstruents, "up to but not including the next stressed vowel" (Beckman 1998: 157, her emphasis.) The examples below demonstrate how this long-distance nasal harmony is stopped by a stressed vowel, whether nasal or oral (note that due to the typographic messiness of marking nasality and stress above vowels, I have marked nasality by underlining each nasalized segment):

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

38) *Guaraní Nasal harmony blocked by stressed vowels* (Beckman 1998: 159, 178; citing Poser 1982: 130; Gregores and Suárez, 1967)

- a) /re + xó + ta + ramó/ → [rexótaramó] 'if you go'
- b) /a^mbaapó + ro + rey + ú/ → [a^mbaapóroreyú] 'if I work you come'¹⁸

To capture the fact that only stressed vowels are freely nasalized, the positional faithfulness story uses the ranking Ident[nasal]-σ' >> *NasalV >> Ident[Nasal]-Seg:

39) *Faithfulness to input nasality*

a) <i>stressed vowels can be nasal...</i>				b) <i>... but unstressed vowels neutralize to oral</i>			
/tupá/	Id[nas]	*NasalV	Id[nas]	/tupá/	Id[nas]	*NasalV	Id[nas]
	σ'		Seg		σ'		Seg
☞ tupá		*		tupá		*!	
tupá	*!		*	☞ tupá			*

The positional faithfulness constraint for stressed vowels will also capture the fact that stressed oral vowels block the continued spread of nasality in data like (38). Whatever constraint drives nasal harmony (here I simply adopt Beckman's use of Align-L[nasal]), this markedness constraint is also ranked between the two Ident[nasal] constraints and so can be violated only to preserve the nasality of a stressed vowel (compare 40ii and iii):

40) *Blocking nasal harmony by stressed vowels (adapted from Beckman 1998:179)*

/re + xó + ta + <u>ramó</u> /	Ident[nasal]-σ'	Align-L(nasal)	Ident[nasal]-Seg
(i) rexó <u>taramó</u>		*****!***	
☞ (ii) rexó <u>taramó</u>		****	****
(iii) rexó <u>taramó</u>	*!		*****

¹⁸Beckman (footnote 10) tells us that the morpheme given in the input of this example as /ro/ with an unstressed nasal vowel is in fact a reduced version of the conjunction/postposition /ramó/ seen in the previous example, whose nasality is expected. There is also some rightward nasal spreading from /ro/s unstressed vowel; see Beckman's footnote 9 on the phonological status and treatment of progressive nasal harmony in Guaraní, which does not concern us here.

The aspect of Guarani that argues crucially for a positional faithfulness account is this pattern of blocking in (40) above. The positional markedness alternative to account for Guarani would replace the constraint Ident[nasal]-σ' with a constraint like License[nasal] – a markedness constraint that requires nasal vowels to be associated with strong positions, such as the stressed syllable (see Flemming 1993; Steriade, 1995 for a fuller story.) The problem for this account is how to handle the mapping in (40) – why should stressed oral vowels block the spread of nasality? As Beckman puts it, the claim of License[nasal] is that “[+nasal] is licensed whenever it is associated to a stressed syllable, regardless of its input source. The underlying nasality/orality of the stressed vowel is irrelevant” (Beckman 1998: 183.) And as (41) below shows, no matter how these markedness constraints are re-ranked they cannot explain why nasal harmony does not spread through a stressed vowel that is underlyingly oral:

41) Failure to block nasal harmony with License[nasal] (from Beckman 1998:183)

/re + xó + ta + ramó/	License[nasal]	Align-L(nasal)	Ident[nasal]-Seg
(i) rexótaramó		*!*****	
⊖ (ii) rexótaramó		*!***	****
⊕ (iii) rexótaramó			*****

This ability to block processes that would change the input specifications of strong positions is thus another reason to maintain the existence of specific faithfulness in the grammar.

The third argument about the best definition of contexts is exemplified by the proposal of Noun-Faithfulness in Smith (2001). Smith demonstrates that a variety of languages display a wider set of contrasts in nouns than other categories (verbs, adjectives, function words, etc.) One of Smith’s examples comes from the accentual

system of Fukuoka dialects of Japanese, in which nouns escape the otherwise-general pattern of pitch accents on the penultimate mora of lexical items:

42) Pitch accent in Fukuoka Japanese:

(a) Verbs/adjectives: pitch accent on σ with penultimate mora:
 tabé̌ta ‘ate’ aká̌ka ~ aká̌i ‘red’
 tabé̌n ‘to eat’ akakaró̌o ‘probably red’

(b) Nouns: faithful to pitch accent on other moras,
 inó̌ti ‘life’
 óokami ‘wolf’ (initial mora accented)

... and to lack of accent on the penultimate mora:
 atama ‘head’ (unaccented)

Smith’s point includes the fact that this special property of nouns appears to be fairly asymmetric: that languages like Fukuoka-prime in which pitch accent is predictably assigned in all morphological categories except verbs, except adjectives, etc. are unattested. Thus, positing markedness constraints relativized to every lexical class except nouns seems to suggest we’re missing something: i.e., faithfulness to nouns.

With these arguments in hand, we will now return to the learning discussion under the assumption that positional faithfulness constraints form part of CON, and so must form part of our learnability story.

4.1.2 Stringency, not fixed rankings

To use the terminology of Prince (1997) and de Lacy (2002): the sets of faithfulness constraints that I have adopted in 35) above stand in *stringency* relations. The constraint Ident-Onset[F] is *less stringent* than Ident-Segment[F] because the former assigns a proper subset of the violation marks assigned by the latter; in other words, violating a less stringent constraint *entails* violating a more stringent one.

The effects of stringency have also been derived using fixed rankings of *two sets* of specific constraints. For example, McCarthy and Prince (1995)'s approach to the phonological privilege of roots is to split faithfulness into Root-Faith and Affix-Faith versions, and to propose the 'meta-ranking' (fixed ranking) of Root >> Affix.

Throughout this dissertation, I will be adopting stringency rather than fixed rankings. For starters, we will see below in section 4.3 that stringency relations between faithfulness constraints can be *language-specific*, and therefore uncapturable in any fixed ranking. And though they are often similar in their effects, the fixed-ranking and stringency approaches also do make different typological predictions. These differences stem from the effects of what Prince terms an 'Anti-Paninian' ranking.¹⁹ An Anti-Paninian ranking is a crucial ranking of a less stringent constraint above more stringent one (i.e., General-F >> Specific-F) – and the effects of Anti-Paninian rankings cannot be replicated in a fixed ranking model, because they are precisely what the model prevents.

While the need for Anti-Paninian rankings between *markedness* constraints, and therefore the use of stringent definitions, seems fairly solid (see especially de Lacy (2002)'s extensive cross-linguistic discussion of sonority-driven stress) the choice between fixed rankings and stringency for faithfulness is less clear. Several proposals

¹⁹ See the Prince and Smolensky (1993) appendix on Panini's theorem, which gives rise to the term.

have been made which rely on General-Faith >> Specific-Faith rankings to produce attested patterns: see Keer (1999: 82-85) on Fula geminate hardening; Lombardi (1999) on Swedish voicing assimilation; de Lacy (2002: chapter 8) on Chipeweyan coalescence and other patterns.²⁰ However these analyses suffer from a typologically-uncomfortable prediction known as 'Majority Rules' (see Bakovic, 1999ab; Lombardi, 1999; Wilson, 2000; de Lacy 2002: §7.7.3) -- a problem whose real scope and possible solution I will not address here. I will return to Anti-Paninian rankings and phonotactic learning in §7.3.

4.2 The learnability argument for a Specific-F >> General-F bias: Smith (2000)

In the same spirit as the high-ranking M bias: we want our learner to assume rankings that resolve errors while being IO-faithful in as few contexts as possible. This means that when choosing a ranking that is faithful to the input, the learner should only install the *least stringent* (i.e. *most specific*) F constraint that can resolve an error. This bias aims at avoiding the learning of superset grammars, as spelled out in Smith (1999, 2000) as well as in Hayes (2004).

Imagine that the learner is confronted with the error in 43):

43) *an ambiguous ERC*

<i>input</i>	<i>winner ~ loser</i>	*mid	Ident[mid]	Ident[mid]-σ1
/bedat/	bedat ~ bidat	L	W	W

In order to resolve this error, the learner knows that *some* faithfulness must be installed above *mid; that is, they can choose between one of these two rankings:

²⁰ See also the rankings in Strujke (2002) between other Faithfulness and her Existential Faith constraints.

44) *Two ways to resolve the ERC in 43)*

- a) Ident(mid)-σ₁ >> *mid >> Ident(mid)
- b) Ident(mid) >> *mid >> Ident(mid)-σ₁

Of these two rankings, 44a) is the more restrictive, because it accounts for the winner's mid vowel while leaving the height of non-initial vowels up to markedness. This grammar is one in which mid vowels are allowed to surface faithfully only in initial syllables; elsewhere they are still ruled out by *mid:

45) *The restrictive results of 44a)*

a) *initial mid vowels survive*

/bedat/	Ident (mid)-σ ₁	*mid	Ident (mid)
(i) \varnothing bedat		*	
(ii) bidat	*!		*

b) *... but non-initial ones do not*

/bedat/	Ident (mid)-σ ₁	*mid	Ident (mid)
(i) badet		*!	
(ii) \varnothing badit			*

Assuming the other ranking in 44b), however, means that *any* input mid vowel will be able to surface faithfully:

46) *The possible overgeneration of 44b):*

a) *initial mid vowels survive*

/bedat/	Ident (mid)	*mid	Ident (mid)-σ ₁
(i) \varnothing bedat		*	
(ii) bidat	*!		

b) *... and so do others!*

/bedat/	Ident (mid)	*mid	Ident (mid)-σ ₁
(i) \varnothing badet		*	
(ii) badit	*!		*

The problem with 46) is the usual superset problem: that if the target language does in fact only permit mid vowels in initial syllables, the grammar in 46) won't cause any further mid vowel errors, and so won't provide any evidence that an overly-

permissive language has been learned. If however the reverse error has been made – the learner has chosen the grammar of 45) instead of 46) – the learner *will* get evidence that they've made the wrong decision. Once they hear a mid vowel in a non-initial syllable, they'll make an error that creates an ERC row like in 47) below. This error clearly demonstrates that installing Ident(mid)-σ₁ will not account for the target's full range of marked vowels, so that Ident(mid) must be used:

47) *an unambiguous ERC row that chooses the ranking in 46)*

input	winner ~ loser	*mid	Ident[mid]	Ident[mid]-σ ₁
/bedat/	badet ~ badit	L	W	e

The same relationship holds true of morphologically-specific faith, e.g.:

48) *an ERC row adapted from Smith (1999)*

input	winner ~ loser	NoCoda	Max(Seg)-Rt	Max(Seg)
/bedat/	bedat ~ beda	L	W	W

When presented with this error, the learner must choose the ranking in 49a) below, and not 49b). This ensures that if the target language only permits codas in roots, the learner will not mistakenly assume that affixes can have codas too.

49) *The two ranking possibilities given 48)*

- a) Max(Seg)-Rt >> NoCoda >> Max(Seg) (restrictive)
- b) Max(Seg) >> NoCoda >> Max(Seg)-Rt (not restrictive)

The conclusion, then, is that finding the most restrictive grammar compatible with a set of ERC rows depends in part on installing the most specific W-assigning faithfulness constraint above each L-assigning markedness constraint.

4.3 The problems of enforcing the Spec-F >> General-F bias

4.3.1 Language-specific relations between faithfulness constraints

As Prince and Tesar (2004) demonstrate, implementing a specific >> general faithfulness bias is not at all straightforward. The previous two biases were easy to enforce because they make reference to language-independent properties. Markedness or Faithfulness, Output-Output or Input-Output – these are definitional properties of a constraint. But there is not always something intrinsic to the definition of a constraint that puts it in a special to general relationship with another – and while many specific to general relations are universal, Prince and Tesar also point out that the stringency relations between faithfulness constraints can be *contingent*. In contingent cases, it is only the interaction of other high-ranking constraints that carve up the space of possible surface forms in such a way to make a particular faithfulness constraint less or more specific than another.

4.3.1.1 Prince and Tesar’s example

Here is the extent of the prolem. Imagine that our learner has constructed the following ERC row:

50) *another ambiguous ERC row*

<i>winner-loser</i>	*[mid]	Ident-mid-σ1	Ident-mid-σ'
képa ~ kipa	L	W	W

Which of the IO-faithfulness constraints that prefer the winner should be installed? What follows is Prince and Tesar’s demonstration that given other facts about the target language (in particular, its pattern of stress assignment), these two faithfulness constraints might stand in *either* stringency relation.

In Language A, the correct generalization is that mid vowels only appear in initial syllables – that is, the true ranking in the language is in 51)

$$51) \quad H_{LA}: \quad \text{Id}(\text{mid})-\sigma_1 \gg *mid \gg \text{Id}(\text{mid})-\sigma'$$

Suppose further that this language assigns stress without fail to the initial syllable of every word; it also assign stresses to later syllables in longer words (this is the case in e.g. Pintupi; see Hayes, 1995 and references therein.) In such a grammar, the initial syllable context is in fact more specific than the stressed syllable context – because of how stress is assigned, every initial syllable is stressed, but not every stressed syllable is initial.

Now consider Language B, whose Support table also includes the entry in 50) but whose correct ranking is the reverse of Language A:

$$52) \quad H_{LB}: \quad \text{Id}(\text{mid})-\sigma' \gg *mid \gg \text{Id}(\text{mid})-\sigma_1$$

In language B, stress is always confined to the initial syllable of a word, but some words do not bear stress at all.²¹ As a result, these two contexts (and associated faithfulness

²¹ On the plausibility of such a language, Prince and Tesar (2004) cite the example of Seneca, (e.g. Michelson, 1988), where “stress behaves more like pitch accent [and] stressless words may occur”. Perhaps

constraints) stand here in the *opposite* specificity relation: every stressed syllable is initial, but not every initial syllable is stressed.

In their search for the most restrictive grammar, the learners of Language A or Language B are in equally dangerous but opposite situations. In resolving this one error, the specificity relations between Ident(mid)-σ1 and Ident(mid)-σ' are crucial to choosing the subset grammar – but depending on the language *either* relation could hold. And the language-specific evidence as to which context is more specific than the other can't be read off any faithfulness constraint to mid vowels, initial syllables or stressed syllables. Instead, these facts are only buried away in the constraint rankings that determine stress – Align-Head-L, Trochee vs. Iamb, and the like.

4.3.1.2 A morphological example

Given the centrality of this issue to my argument, it is worth seeing that this problem is not just a function of initial and stressed syllables. Contingent specificity relations between positional contexts will also emerge as the result of *morphologically*-specific constraints like Root-Faith. For this case, we can use the same ERC, only slightly modified:

53) *a morphologically ambiguous ERC*

<i>input</i>	<i>winner-loser</i>	*[mid]	Ident-mid-σ1	Ident-mid-Root
/kɛpa/	képa ~ kipa	L	W	W

The languages that could have created this error include the following two in which the stringency relation is crucial, but again in either direction. In Language A, what we should take from this is that the constraint in this contingent stringency relationship would really be Ident[mid]-PitchAccented-σ.

there are no prefixes or pro-cliticizing elements, so every initial syllable is also a root syllable. But the reverse is not true, since roots can be bigger than a syllable. So in language A: initial syllables are a special case of root syllables.

In the second language, affairs are different. Language B again has no prefixes, but its roots are small – in fact, no longer than a syllable – meaning that root syllables are always initial syllables. Furthermore, this language has free-standing words that do not count as roots – i.e., function words can create their own Prosodic Words, so some initial syllables are not root syllables. Thus in language B, root syllables are a special case of initial syllables.

The resulting learning problem is just as in the previous section. Language A could restrict mid vowels only to initial syllables, and Language B could restrict mid vowels only to *roots* – both meaning that only the Ident constraint referring to the more specific context should be installed above *[mid]. But how can the learner know which Ident constraint that is? Again, the fact that root syllables are more or less specific than initial syllables is encoded only in the ranking of constraints that say nothing about the mid vowel that caused the error in 53), but rather in the ranking of constraints that e.g. align morphological roots and prosodic words (McCarthy and Prince, 1993 *et seq.*)

4.4 Interim Summary

In the face of the problem raised above, Prince and Tesar (2004) go so far as to suggest that positional faithfulness in fact be barred from CON, and replaced with positional markedness constraints instead (e.g. Zoll, 1998) whose specificity relations pose no problem for learning, as discussed in §3.3. Nevertheless, the claim of this

dissertation is that the kinds of evidence marshaled at the beginning of this discussion (§4.1.1) require us to include positional faithfulness constraint in our typology, and therefore that their consequences for restrictiveness must be accommodated by our learner.²² And since the previous sections have demonstrated that a restrictive learner must be biased to choose the faithfulness constraint possible with the most specific context when installing W-preferring constraints: our learner will have to be able to discover these relations.

Returning to the broader picture: sections 2 through 4 have presented the Biased Constraint Demotion approach to restrictiveness, and the three ranking biases that I adopt in my version of this algorithm. Before moving on, the next section returns to the role of the Support in BCD learning, and uses the biases we've now seen to emphasize its key role in the on-going quest for restrictiveness. Then, in section 6, I will return to the proper treatment of the specific \gg general faithfulness bias, given the problems just raised in §4.3.

5. Returning to the role of the Support

As was emphasized at the beginning of this chapter, the Rich Base assumed in Optimality Theory means that language-specific knowledge in an adult OT grammar is instantiated fully in rankings, rather than in lexical items. Nevertheless, in the BCD approach it is in fact the Support and *not* the constraint rankings that are the real locus of learning over time. BCD is a function from Support data to a ranking; the BCD learner is gradual and incremental at the level of the Support, but quick and flexible in its rankings.

²² See Beckman, Jessen and Ringen (2006) for an interesting different kind of argument for positional faith over positional markedness.

The calculation from the Support to a ranking is something the learner can do any time it wants, and there's nothing sacrosanct about the current ranking per se. Learning is a continual process of updating the Support, and then seeing what rankings that Support permits the BCD algorithm to construct.

As a relevant comparison: Ito and Mester (1999) propose a learning strategy called Ranking Conservatism, which implements the $M \gg F$ ranking at the initial state and also throughout learning, though in a formally rather different way than does BCD. Roughly speaking, the Ranking Conservative learner proceeds gradually away from the initial state by re-ranking in ways that account for the errors made but otherwise keep as many M constraints above as many F constraints as possible. As it turns out, this imperative is not enough to ensure that learners do not adopt superset grammars by accident: Prince and Tesar (2004) discuss two cases in their Appendix A in which Ranking Conservatism isn't enough; see also Broiher (1995).

5.1 A kind of learning error: winner misparses

The BCD's reliance on the Support also makes it resilient to superset traps created by missing or incorrect assumptions about hidden structure. Since the learner is never committed to its rankings independent of its Support: if new information comes to light in the Support it will be reflected in the ranking. The kinds of mistaken hidden structure that could in principle get into the Support are varied: wrong morphological categories or decomposition, wrong syllabification or footing, or similar. I will call these mistakes 'winner misparses'.

The first point about such misparses is that, like other errors, they are all a function of the current ranking. An upcoming example is the syllabification of a word-medial, post-tonic cluster like [kábla]. On the one hand, NoCoda prefers the complex onset parse, as in (51a). On the other hand, however, the constraint Stress-to-Weight (e.g. Hanson and Kiparsky 1996; Elenbaas, 1999; Elenbaas and Kager, 1999) prefers syllabifying this cluster as a coda-onset cluster, because it requires that stressed syllables be closed. What will decide between these syllabifications is thus relative re-ranking of these two constraints at the point when the learner hears this winner:

54) *Choosing the syllabification of the winner*

(a)			(b)		
[kábla]	NoCoda	Stress-to-Weight	[kábla]	Stress-to-Weight	NoCoda
☞ ká.bla		*	ká.bla	*!	
káb.la	*!		☞ káb.la		*

Note that since they are based on constraint rankings, winner misparses are not necessarily infrequent – that is, they are not one-time glitches. If the learner’s current grammar assigns the wrong syllable structure or other representation to a class of winners, they will continue to make these parsing errors until some further learning takes place.

The point here that is these winner misparses can prevent the correct acquisition of *other* aspects of the grammar, and that overcoming their influence requires remembering errors so one can undo the rankings that winner misparses caused. This is what the Support allows us to do.

5.2 **How the Support allows BCD to overcome winner misparses**

Imagine the learner is acquiring a language with coda devoicing, where onset obstruents can be voiced or voiceless but codas are always voiceless.²³ This grammar requires the simple ranking in 48) below:

55) *The target grammar*
Ident[voice]-Onset >> *VoicedObs >> Ident[voice]

One thing the learner must do to learn this grammar is to correctly syllabify all voiced obstruents as onsets. If it does this correctly, then its ERC rows will look like 53) below:

56) *Learning onset voicing: the right winner parse*

winner ~ loser	*VoicedObs	Ident[voice]-Ons	Ident[voice]
[ká.bla] ~ [ká.pla]	L	W	W

As we saw in 4.2, a bias for the most specific faithfulness constraint will ensure that this ERC row will teach the BCD learner the right ranking in 56). The problem illustrated in 54), however, is that an incorrect ranking of the constraints Stress-to-Weight and NoCoda could drive this learner to the coda-onset syllabification of this cluster. If the learner has adopted the 54b) ranking in which Stress-to-Weight chooses the coda-onset sequence, this will change the violations in their ERC row for kábla ~ kapla. In this grammar, the winner’s [b] and loser’s [p] of this cluster are syllabified as a coda, so Ident[voice]-Onset will not make a choice between them:

²³ Thanks to John McCarthy for suggesting this example.

57) *The misparsed ERC row*

winner ~ loser	*VoicedObs	Ident[voice]-Ons	Ident[voice]
[káb.la ~ káp.la]	L	e	W

The only grammar that BCD can learn from this ERC row is the ranking below:

58) *The superset grammar learned from 57):*

Ident[voice]-Ons >> **Ident[voice]** >> ***VoicedObstruent**

And this ranking defines a superset language compared to 55), because it allows a spurious voicing contrast in coda position:

59) *The restrictiveness problem with 58):*

<i>hypothetical</i> /káb/	Ident[voice]-Ons	Ident[voice]	*VoicedObs
⊕ káb			*
káp		*!	

So for the present, the learner has acquired a superset grammar – so long as they have the learning ERC row of 57) in their Support, every cycle of BCD re-ranking will generate a grammar with general Ident[voice] ranked too high.

But this error is not a permanent overgeneralization for the learner. When the learner gets evidence for the ranking NoCoda >> Stress-to-Weight, he or she will now have a way to update his or her Support entries – fixing the input and winner representations, and thereby calculating the correct constraint violations as in 56). (I do not go into the details here of how such further learning operates – see Tesar and Smolensky 2000's notion of RIP and the acquisition of hidden structure like footing; see Tesar et al 2003 for their discussion of a proposal about later morphological reparing of

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

winners in the Support, using the process called Surgery.) But once it has happened, the first time they re-rank *after* that Support update, they will now choose the right ranking.²⁴

Note that the learning cycle that gets the learner to the correct grammar will *not* come as a result of any voicing errors. Once BCD has chosen a grammar on the basis of the incorrect parse, it now allows voiced obstruents in all syllabic positions, so it cannot make any errors in phonotactic learning. The reason the BCD learner can nevertheless overcome these winner misparses, and revert to a subset grammar is that it stores the errors it is no longer making. When its ranking changes in a way that can affect old errors (e.g. re-syllabify them), it can re-calculate the relevant constraint violations and so use them to choose the newest, most restrictive grammar.

5.3 A second example²⁵

This example involves a different kind of winner misparse which has received considerable discussion: the footing of a trisyllabic word with medial stress, as either trochaic [σ (σ σ)] or [(σ σ) σ] (see Tesar, 2000.)

Imagine our learner is now acquiring an iambic language with foot-initial strengthening: foot-initial stops must be aspirated, but all others are unaspirated. This language requires has the allophonic ranking in 60) below:

60) *The target grammar*

Foot-Initial Aspiration >> ***Aspirate** >> **Ident-[laryngeal]**

²⁴ It is worth noting that choosing the right syllabification of a medial consonant sequence is far from easy. It may indeed depend on something quite subtle process that e.g. can be understood as triggered only by closed syllables, and which is *not* triggered in the first syllable of CVbIV words with initial stress.

²⁵ Thanks to Joe Pater for suggesting this example.

To correctly diagnose this allophonic ranking, the learner must again have the winner's representations correct. If the learner has parsed all its feet into iambs, then its ERCs rows will look like 61), and BCD will be able to choose the right ranking.

61) *Learning foot-initial aspiration: the right winner parse*

/p ^h abóla/	*Aspirate	Ft-InitialAsp	Ident[laryng]
[(p ^h abó)la] ~ [(pabó)la]	L	W	W

Imagine however that the learner were to make the error above early on, at a point when foot form had not yet been decided. What will decide between these foot structures is the relative ranking of Trochee and Iamb:

62) *Choosing the footing of the winner*

(a)			(b)		
/p ^h abóla/	Iamb	Trochee	/p ^h abóla/	Trochee	Iamb
☞ (p ^h abó)la		*	(p ^h abó)la	*!	
p ^h a(bóla)	*!		☞ p ^h a(bóla)		*

If the learner has adopted the ranking in (b) instead of (a) (even just for this error), this will create a different ERC row, precisely with respect to the foot-initial aspiration constraint:

63) *The misparsed ERC row*

/p ^h abóla/	*Aspirate	Ft-InitialAsp	Ident[laryng]
[p ^h a(bóla)] ~ [pa(bóla)]	L	e	W

And the only grammar that BCD can learn from 63) is the superset one below:

64) *The superset grammar:*
Ft-InitialAsp >> **Ident[laryngeal]** >> ***Aspirate**

This ranking defines a superset language compared to the target in 60), because it allows a spurious laryngeal contrast outside the foot-initial position:

65) *The restrictiveness problem:*

<i>hypothetical</i>	Ft-InitialAsp	Ident[laryng]	*Aspirate
/bop ^h a/			
☞ (bop ^h a)			*
⊗ (bopa)		*!	

Again: this winner misparse has led our learner to acquire a superset grammar. But once Trochee has been properly re-ranked above Iamb, the learner can re-evaluate their ERC rows like 61) to look instead like 63), which will finally lead to the right allophonic ranking.

5.4 **Summary**

This section has highlighted a crucial role of stored errors in biased learning – the escape from superset grammars when early learning data is re-interpreted. I will also return to this point in later chapters, when I demonstrate the trouble that a memory-less learner like the GLA has in keeping restrictive despite winner misparses (chapter 3 §4; chapter 4 §7.3.2.)

With the centrality of the Support fully in mind, we can now return to the problem of the specific-F ranking bias, and see how learners can use their current Support to handle even contingent F-subset relations.

6. The proposal: finding the most specific IO-Faith constraint

Section 4 provided arguments (i) that our theory should contain positional IO-faithfulness constraints, (ii) that their presence in CON requires a bias for BCD to rank the most specific constraints as high as possible, and (iii) that specific-to-general relations between faithfulness constraints are not all universal, and can differ crucially from language to language. With all these claims in mind, this section presents the proposal for properly constructing this bias.

6.1 The goal: determining subset relations between the contexts of faith

To adequately determine specific-to-general faithfulness relations, the learner must somehow access information about the rest of the language being learned, which our BCD learner is storing in the Support. Central to what follows is the idea that to discover the specificity of faithfulness constraints, learners must abstract away from particular constraints and instead consider the *contexts* of those constraints. This move is somewhat subtle, but it represents a rather different approach to the problem than the one envisioned in Prince and Tesar (2004)'s discussion, so I will endeavour to clarify the matter below.

First, the problem. If a language has a crucial ranking of a specific faithfulness constraint above a more general one, it is because something marked only appears in the specific position but it is banned in the general one. What this means for the learner is that in errors during phonotactic learning, the constraint violations assigned by the specific and general IO-Faith constraints will be identical: when either assigns a W, they both will:

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

66) *specific-F = general F, in the subset grammar's ERC rows*

<i>input</i>	<i>winner ~ loser</i>	*mid	Ident[mid]-Seg	Ident[mid]-σ1
/bedat/	bedat ~ bidat	L	W	W

No errors will exist in which Ident(mid) assigns a W but Ident(mid)-σ₁ does not. Thus, our learner is not going to learn that Ident[mid]-σ₁ is more specific than Ident[mid] by examining this ERC row – something must be abstracted away from.

In my proposal, the learner's attention is directed away from the Ws assigned by the faithfulness constraints in 66), and even away from mid vowels, and instead aimed at the contexts "Seg" and "σ₁". The learner will examine the observed winners of their language, and determine whether these positional contexts sit in a subset/superset relation or not; from that information they will choose constraints to install.

6.1.1 Constraint stringency vs. context specificity

The way of calculating the specific-to-general bias that I will propose relies on context specificity, rather than constraint stringency itself, to guide the learner's ranking decisions.²⁶ It is important to see that the specificity of *contexts* does not translate straight to the stringency of *constraints* – because faithfulness constraints have both contexts and also banned mappings. In 66) above these two properties do line up, because the two faithfulness constraints at hand are both relative to the vowel feature [mid]. Thus we can say both that the context 'σ₁' is more specific than "Seg", and also that the Ident[mid]-σ₁ constraint is less stringent than Ident[mid]-Seg.

²⁶ The method Hayes (2004) uses to impose his Favour Specificity bias also uses context specificity rather than constraint stringency; see section 7.2 for more details.

But stringency relations do not hold between constraints that protect different features. For example, Ident[**mid**]-σ1 is not less stringent than Ident[**voice**]-Seg because they penalize completely different mappings, so each can be violated independent of the other:

67) *Context specificity, but no constraint stringency*

/peg/	Ident[mid]-σ1	Ident[voice]-Seg
peg		
pig	*	
pek		*
pik	*	*

I will argue in section 6.5 below that there is good reason to focus on contexts rather than constraints – because even when abstracting away from the unenlightening cases like 66), constraint violations can be misleading when it comes to *contingent* stringency relationships. To understand the proposal, however, it is enough to remember that this search for subset/superset relations will be focused on contexts, and only then will apply the search’s findings to faithfulness constraints and their ranking.

6.1.2 Outline of the proposal

To summarize so far: what we need is a way to detect any specificity relationship between any two faithfulness *contexts* C1 and C2. Broadly speaking, this search is going to involve looking at each instance of C1 among the language’s winners and seeing whether it is *also* an instance of C2, and then vice versa. If the two contexts *do not* stand in a specific-to-general relation, we will only need to come across the two relevant pieces of evidence to determine no such relation exists. To determine that there *is* such a

relation, however, our search will have to continue until all winners have been examined – to be sure that *every* known instance of C1 is also a case of C2.

As we saw in section 4, central to Prince and Tesar’s skepticism about the specific faithfulness bias is the existence of contingent faithfulness stringency – in the present terms, contexts can sit in contingent specificity relations. However, many contexts *always* stand in specificity relation, and it could well be argued that searching the entire Support for evidence of such relations is rather inefficient. One universal relation is simply that the most general faithfulness context – what I have been calling simply “Seg” – is less specific than any faithfulness constraint above the segmental level, which references either a prosodic or morphological category:

68) *Examples of the two kinds of universal prosodic context specificity*

Id(mid)-Onset, Id(mid)-σ ₁ ...	Id(round)-Rt, Id(round)-Noun ...
<i>are less stringent than</i>	<i>are less stringent than</i>
Id(mid)-segment	Id(round)-Segment

The same point can be made at the featural level – that the context “Seg” is more general than any more specific combination of features (e.g. Ident(voice)-Labial, Ident(voice)-Labial and Dorsal.) In fact a stronger claim can be made at the featural level, by adopting the fairly standard OT assumption that there are no language-specific meanings to featural combinations – i.e. that the context “Labial and Dorsal” refers to the same set of segments in every language.²⁷ This means that the relationships between featural contexts is definitional and language-independent, and so *all* their subset and

²⁷ For the alternative type of view, see e.g. Rice and Avery 1989, 1991.

superset relations can merely be read off their constraint definitions. This is illustrated in 69) below:

69) *Two examples of universal featural context specificity*²⁸

IdentPlace-(Stop or Fricative)	IdentVoice-(Nasal)
<i>is less stringent than</i>	<i>is less stringent than</i>
IdentPlace-(Stop or Fricative or Nasal)	IdentVoice-(Nasal or Oral Stop)

As a result of these universals, my approach to calculating IO-faith specificity has two steps. When considering a set of faithfulness constraints, the learner will first ‘pre-compile’ the specificity between their contexts using universal properties of CON – defined below as the kind and number of ‘contextual arguments’ each takes (explanation to follow.) If this first pass underdetermines the specific-to-general relations (as it will in the case of e.g. Ident(F)- σ ’ and Ident(F)- σ_1) – the learner will then turn to the Support’s winners, and evaluate contexts one by one. This second step will require a tool over which the learner can calculate these relationships, which I will call here the Context Table. To summarize:

70) *The method for detecting IO-faith Specificity*

- (a) Is there a universal relation between contexts X and Y?
(the *Context Arguments* test)
- (b) Does the Support reveal a contingent relation between contexts X and Y?
(the *Contingent Specificity* test)

²⁸ For work which uses featural faithfulness constraints of this sort see e.g. Becker (2006) on (OO) Ident(vce)-Labial; Beckman and Ringen (2006) on Ident(vce)-Fricative.

In the sections that I follow I demonstrate how and why these two steps work; integrating them into a BCD-style ranking bias will be the province of section 7.

Before continuing, a remark about efficiency and computation may be in order. Given that the first step of this method establishes contextual relations that are universal, it might be something of a pedantic pity to re-calculate whether e.g. segments that are onsets are a special case of segments every time the learner needs to choose the ranking Ident[F]-Onset >> M >> Ident[F]-Seg. As an alternative, one could either hardwire the learner with the knowledge of these context subset relations, or else ask the learner to apply step 1 every time they need to but then store the results to be used in all subsequent rankings.

If these two alternative approaches were to be implemented computationally, it seems likely that a constant recalculation of universal context specificity would add unnecessary effort to the learner’s task. The only reason to do so would seem to be a purely formal interest in hard-wiring as little into the learner as possible – beyond this aesthetic concern, I leave the matter open.²⁹

6.2 The first step: finding universal specificity relations

The schematics in 68) and 69) showed the two kinds of universal specific-to-general relations we want our learner to notice just from constraint definitions, or rather the definition of their contexts.

²⁹ One substantive issue with hardwiring these universal specificity relations might arise when comparing constraints with multiple contexts, some but not all of which are in universal stringency relations, and thereby deciding e.g. whether Ident[mid]-Root-Seg and Ident[mid]- σ_1 -Labial are in a stringency relationship. Whether the relevant constraints are ever in conflict in a ranking situation is an empirical question to which I do not know the answer.

First, we must decide what a context is. All work on positional faithfulness is concerned with giving constraints the right structural context – input initial syllable, output onset, and so forth – as well as its banned unfaithful mapping (voicing mismatch, missing correspondent, etc.) In this work, the contextual arguments of constraints are always at the level of the segment, or above: either a prosodic or morphological category. Some representative examples of constraints and their context arguments are given below:

71)	<i>constraint</i>	<i>positional context argument(s)</i>
(a)	MAX	segment
(b)	DEP- $\{V\}$	vowel & segment
(c)	IDENT[voice]	segment
(d)	IO-IDENT[lab]-Ons	onset & segment
(e)	IO-IDENT[phar]-Rt	root & segment
(f)	IO-IDENT[mid]-Rt- σ 1	initial syllable & root & segment ³⁰

As 71) shows, all the positional constraints have been defined with the argument *segment* as well as something else.³¹ This is important for one thing because it allows the comparison of phonological and morphological contexts, whose affiliations only overlap at the segmental level. With the arguments defined this way, we can find the most specific members of a set of positional contexts merely by finding proper *subsets*:

³⁰ I only include (g) just to show that in principle a constraint could have three context arguments.

³¹ This means I have nothing to say about faithfulness to floating features or their kin – on such objects in the OT context, see e.g. Zoll (1996), Wolf (2005).

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

- 72) *the Positional Context Arguments test*
 GIVEN: two positional context arguments, P1 and P2
 IF: if P1 is a *proper subset* of P2 (that is: every member of P1 is also a member of P2, but some member of P2 is not a member of P1)
 THEN: P1 is *less specific* than P2

By the test in 72), the constraints in 71) include three specificity relations, that are clearly the right ones:

- 73) (i) MAX (a)'s context is less specific than those of constraints (b),(d)-(f)
 (ii) IDENT (c)'s context is less specific than those of constraints (b),(d)-(f)
 (iii) IO-Ident-Rt (e)'s context is less specific than that of IO-Ident-Rt- σ 1 (f)

I am also assuming that faithfulness constraints can be relativized to featural contexts. Thus, we will need to determine specificity relations among the subsegmental context arguments of constraints. One current view of featural faithfulness comes from the theory of de Lacy (2002), in which faithfulness constraints protect features in direct proportion to their markedness, and where featural markedness directly reflects stringent markedness scales. As an example, the featural markedness scale on major places of articulation in 70) below results in a set of IDENT[place] constraints, which increase in stringency from the most specific to the most general:

- 74) *Place of Articulation Markedness Scale*, from most to least marked:
 (scale and constraints adapted from de Lacy, 2002: 167,173-174; see e.g. Jakobson 1941/1967; Paradis and Prunet, 1991; Lombardi, 1995)

dorsal > labial > coronal > glottal

- 75) *constraint examples* *subsegmental context argument(s)*
- (a) IDENT[place]-dors dorsal
- (b) IDENT[place]-dors, lab dorsal or labial
- (c) IDENT[place]-dors, lab, cor dorsal or labial, or coronal
- (d) IDENT[place] dorsal or labial or coronal or glottal

This theory's faithfulness constraints wear their stringency relations on their sleeves. But notice that the relationship between the number of arguments and the specificity of a constraint is opposite to the prosodic domain: here, each argument *adds* another featural context for faithfulness to apply to, rather than further restricting its application. As such, the test for subsegmental arguments equates decreased specificity with *supersets* rather than subsets:

- 76) *the Subsegmental Context Arguments test*
- GIVEN: two subsegmental context arguments, S1 and S2
- IF: if S1 is a *proper superset* of S2
- THEN: S1 is *less specific* than S2

By this reckoning, we can calculate that 75)d)'s context is less specific than all three other constraints; that 75)c)'s context is less specific than a) and b)'s, and so on.

6.3 The second step: finding contingent specificity relations

We now have seen the straightforward way of determining the specific-to-general relations inherent to constraint definitions. In this section I turn to the method of determining contingent specific-to-general relations, using a tool I will call the Context Table (CT). Note that the method for building such tables, though to a slightly different end, was also proposed by Hayes (2004), footnote 31, and that my CTs were (re)designed with his proposal in mind.

A Context Table is a chart that keeps track of whether a faithfulness context can occur independent of another, in a given set of words:

- 77) *A Context Table, not yet filled in*

	σ'	<i>onset</i>
σ'		?
<i>onset</i>	?	

At the left of each row and the top of each column are a series of contexts. In a completed CT, the two cells marked “?” will either be filled in with a check or left blank. To complete the table, the learner must find every instance of each context in every winner. For illustration purposes, imagine the learner has added only three ERC rows to the Support thus far, and that the resulting winners are as in 78):

- 78) *A set of stored words:*
- (a) páť (b) ibák (c) páda

Here is how a context table is completed. To fill in each cell, the learner asks whether each instance of phonological material in the context of row x is also in the context of column y:

- 79) *The Context Table procedure*
- GIVEN: For a set of winners, and a context table with rows r_1 to r_x , columns c_1 to c_y and cells $\langle \text{row}, \text{column} \rangle$
- IF: some segment is in the context r_i but not in the context of c_j
- THEN: put a mark in cell $\langle i, j \rangle$

In 77), there are two contexts, so the ‘if’ statement of 79) above ranges over two questions: “is every segment in a stressed syllable also in an onset?” and “is every segment in an onset also in a stressed syllable?” Given the first question: the first segment in a stressed syllable that is not an onset (say, the á of [pat] in 78a) will cause the learner to update this context table as in 80):

80) *A Context Table, half filled in*

	σ'	<i>onset</i>
σ'		✓
<i>onset</i>	?	

To answer the second question: once the learner comes across the onset of an unstressed syllable (like the [d] of 78c), they will add a second mark to their table:

81) *A Context Table, fully filled in*

	σ'	<i>onset</i>
σ'		✓
<i>onset</i>	✓	

Reading context tables to answer the specificity question works as follows:

82) *The Contingent Specificity test*

GIVEN: two context arguments, C1 and C2, and a context table with cells <row, column>

IF: cell <C1,C2> of a Context Table has a mark while cell <C2,C1> does not

THEN: C1 is *less specific* than C2

As soon as the learner has put marks in cells <C1,C2> and <C2,C1> of a table, the test above already tells them that no stringency relation holds between two constraints – the

search is over. However, if at the end of the search these two cells are asymmetric (one has a mark, and the other doesn't) the learner will have found a *contingent* specificity relation.

To illustrate this, imagine we provide our learner with a novel language, which has the following two attested properties. First: as in Pintupi discussed in §4.3.1.1, stress is assigned using trochaic feet built L-to-R, meaning that initial syllables are always stressed, as well as every odd syllable after that. This means that initial syllables are a special case of stressed syllables. Second, mid vowels in this language are restricted to *initial* syllables only – this is the case in Shona: see esp. Beckman (1998) and references cited therein. So let's call this language Shontupi. With respect to the distribution of mid vowels, the necessary Shontupi ranking is:

83) Ident[mid] $_{\sigma 1}$ >> *mid >> Ident[mid] $_{\sigma}$ (...and Ident[mid] $_{seg}$, etc.)

And in the course of acquiring Shontupi, our learner encounters the by-now familiar ERC row repeated in 84):

84)

/képa/	*[mid]	Ident [mid] $_{\sigma 1}$	Ident-[mid] $_{\sigma}$
képa ~ kípa	L	W	W

Based on this error, our learner can build a context table. To do so, we must have some winners to examine, so:

85) *A representative set of Shontupi words*

képa	típu
kópilá	tábukida

In examining these words using the reasoning from the context table procedure above, the learner will discover e.g. that the segments [la] of 'kópilá' are in the context [σ'] but not [σ_1]. This means that it will enter a mark in cell $\langle \sigma', \sigma_1 \rangle$:

86) *The resulting context table*

	σ_1	σ'
σ_1		
σ'	✓	

On the other hand, it will never come across a word in which some segment is in an initial syllable, but not a stressed syllable. Thus the table in 86) will be its final context table, to which it will apply the Contingent Stringency test:

87) *The Contingent Stringency test, applied to 86)*

SINCE: cell $\langle \sigma', \sigma_1 \rangle$ cell has a mark while cell $\langle \sigma_1, \sigma' \rangle$ does not
THEN: context [σ'] is *less specific* than the context [σ_1]

And with this knowledge, the specific \gg general faith bias our learner will be armed with in section 7 will correctly resolve errors like the one in 84) by installing just the more specific constraint, Ident[mid]- σ_1 .

6.4 Why context tables are dynamic

The context tables that I have proposed here are *dynamic*. They are built on-the-fly – constructed as biases demands, used once to choose between a set of Faithfulness constraints at one particular stratum and then forgotten, and are not stored in any learning memory. In other words, context tables are NOT like the Support, in that they are not the

learner's gradual lexicon of contextual asymmetries encountered so far, built up incrementally over time. This section demonstrates the reason to build CTs dynamically rather than incrementally: to prevent redundancy in the recovery from winner misparses that will add *extra marks* to early context tables.

6.4.1 What can go wrong in a context table?

Two things can go wrong in the building of a context table. Compared to a hypothetical correct table, either a missing mark can be absent from a cell where one should be (89a), or an extra mark can be present in a cell where one shouldn't be (89b):

88) *The correct Context Table for a language*

	σ_1	σ'
σ_1		
σ'	✓	

89) a) *A missing mark*

	σ_1	σ'
σ_1		
σ'		

b) *An extra mark*

	σ_1	σ'
σ_1		✓
σ'	✓	

When will these mistakes get made, and what kinds of restrictiveness problems can they cause?

First: errors of missing marks are easy to make, but they don't cause any permanent restrictiveness problems on either the dynamic or incremental approaches. For example, the learner of Shontupi who hasn't seen any words longer than two syllables yet will only have seen words with one initial and one stressed syllable – being the exact same syllable in each word. If at this point they build a initial syllable/stressed syllable

context table, they will build the one in 89a), missing a mark, and so they will have no stringency reason not to install e.g. Ident-[mid] σ' , which can unfortunately build them a superset grammar.

However, this missing mark will appear as soon the learner encounters a word like e.g. ḱepil̀a, in which the learner can see that stressed syllables need not be initial. And regardless of whether the context table that encodes this discovery is being built from scratch when required by the BCD algorithm, or being augmented incrementally, the next time the learner uses the Context Table procedure it *will* correctly include the mark and so require the correct installation of only Ident-[mid]- σ .

Unlike missing marks, extra marks will get into a context table not by having insufficient data but through the structural misanalyses that I have called ‘winner misparses’. In the present case, a relevant winner misparse would make the learner of Shontuṕi incorrectly believe they’d heard an unstressed initial syllable. This could result from misparsing a two-word sequence (clitic-noun, preposition-verb etc.) as one word, making it appear that stress falls on the second syllable of this false ‘word’:

- 90) *A winner misparse*
 a) correct parse: [ba] [ḱepil̀a]
 b) potential misparse: [baḱepil̀a]

This misparse will result in a CT with an extra mark in the top right-hand corner:

- 91) *the extra mark caused by a winner misparse*

	σl	σ'
σl		✓ (k ́)
σ'	✓	

Like with the missing mark, this CT wrongly assures the learner that there is no stringency relationship between initial and stressed syllable contexts, and so creates the possibility that Ident- σ' constraints will be installed when Ident- σ_1 constraints should be used.

As I argued in previous sections, winner misparses are a likely part of the learning process, because getting the right structural analyses of words depends on grammatical and lexical properties that are not available anywhere in the acoustic signal. Another such example where a winner misparse would choose a superset grammar comes from a language in which stressed syllables are always in the root. If the learner of this language misparses a word’s root-affix boundary as in 92b) below, the apparent fact that stress can appear on an affix would cause the creation of the incorrect context table in 93b):

- 92) *A morphological misparse*
 a) correct parse: $[[\sigma' \sigma \sigma]_{\text{root}}]_{\text{word}}$
 b) potential misparse: $[\sigma'_{\text{affix}} [\sigma \sigma]_{\text{root}}]_{\text{word}}$

- 93) a) *the correct CT*
- | | | |
|-------------|-------------|-----------|
| | root | σ' |
| root | | ✓ |
| σ' | | |
- b) *the extra mark CT*
- | | | |
|-------------|-------------|-----------|
| | root | σ' |
| root | | ✓ |
| σ' | ✓ | |

(The *other* mark, which registers the fact that root syllables need not also be stressed syllables, will be correctly drawn in response to either parse – because in both winners the second and third syllables are in the root but unstressed.)

6.4.2 Overcoming extra marks in a context table

Under the dynamic approach to context tables, extra marks are just hiccoughs in data processing – in the same way that winner misparses are themselves. If the Shontupi learner draws a misleading CT because of a bad parse, it may cause the temporary construction of a superset grammar, but this error will only last as long as the bad parse does. Once a word is reparsed and its Support entry has been corrected (e.g. 93b has been replaced with 93a), the *next* cycle of learning will require a new CT to be drawn and the right stringency relations will emerge.

In an incremental approach, however, extra marks would be somewhat messier to overcome, because they'd require a separate clean-up strategy in the CT-building mechanism. Once the learner had realized that [ba képilà] was in fact two words, she would have to entertain the re-calculation of *every marked cell* one of whose structural contexts was in the re-parsed word.³²

Furthermore: making context tables dynamic leaves the Support as the one true and constant repository of learning data, in keeping with (at least part of) the BCD spirit. As the Support grows and changes, its emerging knowledge will influence both the frequent re-ranking of constraints and the frequent recalculation of dynamic context tables, both when prompted by the learning algorithm.

³² This clean-up strategy process would work roughly as follows: for every cell <x,y> that has a mark, determine whether the re-parsed word provides evidence for that mark, or a mark already in the mirror image cell <y,x>. If it does, move on to the next marked cell. If it doesn't, however – this could be because the winner misparse was the cause of this mark, and it should be removed, or because *some other* word in the lexicon put it here – so the learner must re-determine whether this cell should have a mark using *all* the entries in the Support. In other words – they must re-build this cell from scratch precisely as the dynamic CT is built.

6.5 Why contingent specificity cannot be learned from Ls and Ws

In this section I demonstrate why contingent specificity relations cannot come from the comparison of the ERC rows themselves – that is, why they cannot be assessed by comparing the behaviour of faithfulness *constraints* rather than *contexts* (recall the discussion in section 6.1 above on the difference.)

What the learner wants to know about the target language is whether faithfulness in one of the two contexts under consideration *implies* faithfulness in the other, but not vice versa. So with the contexts of initial and stressed syllables in mind, one approach might have been to look across the Support for winner-loser pairs in which faithfulness to some feature is decisive in one of the two domains (assigning a W), but ambivalent in the other (assigning an e).

To give a concrete example that can demonstrate this approach's failure, let us continue to imagine that our target language is Pintupi in which initial syllables are a special case of stressed syllables. So in looking for asymmetric ERC rows as suggested in the previous paragraph, the learner will be hoping for examples like 94) below:

94) An asymmetric ERC row

/kípy/	*front+rd	Ident-hi- σ_1	Ident-hi- σ'
kípy ~ kipi	L	e	W

In this approach, this ERC would act as a hint that Ident(hi)- σ_1 is a more stringent constraint in the target language than Ident(hi)- σ' . This in turn would translate into a specificity relation between the two contexts [σ_1] and [σ'].

As it turns out, however, the nature of positional faithfulness makes this kind of evidence from ERC rows unreliable – even contradictory. Because when searching the

Support for asymmetric rows like 94), an unlucky learner of Pintupi might make the error in 95a), and build the ERC in 95b) as a result:

95)a) *An unfortunate error:*

/kýpa/	*front+rd	Ident(hi)- σ_1	Ident(hi)- σ'
kýpa	*		
~ kipá		*	

95)b) *The resulting perverse ERC*

/kýpa/	*front+rd	Ident(hi)- σ_1	Ident(hi)- σ'
kýpa ~ kipá	L	W	e

In this language, this ERC demonstrates the *reverse* asymmetry compared to 94). As such, it is completely misleading to the learner: it suggests that stressed syllables are a special case of initial syllables! As we know that the target language's initial syllables are always stressed, this can't be right.

The crucial problem with using an ERC row like 95b) to reason in the way we did previously with 94) comes from the differences between their winners and losers. In 95b), the loser differs from the winner in its violation of Ident(hi)- σ_1 , because it has unrounded the initial syllable. However, this loser also differs from the winner by shifting stress onto the second syllable. And it is because of this second change that the loser does not violate Ident(hi)- σ' : not because the input's stressed syllable height has been preserved, but because the input's stress has been moved. This problem is in some sense inherent to a pathology in the definition of positional faithfulness – one which has been raised by a number of authors³³ and which remains unresolved.

³³ See Beckman (1998) citing Rolf Noyer, as well as Wilson (2000).

7. Implementing the Spec-F >> Gen-F bias

7.1 A working BCD algorithm

With all the results of the chapter so far, we are now in a position to put together our rankings biases into a working BCD algorithm. To understand how this works, we need one more piece of Prince and Tesar's proposal; I will describe it here rather briefly, but the reader who is unfamiliar with BCD is encouraged to consult the much more thorough explication of these ideas in Prince and Tesar's work.

We have already seen that ranking biases drive the learner to install e.g. markedness higher than faithfulness *when possible*. To get from ranking biases to a BCD algorithm, we must see how the learner stays as close to their biases as possible even when the data makes installing any of the preferred constraints *impossible*.

The leading idea from Prince and Tesar on this aspect of the algorithm is that the learner should install just as many of the dispreferred constraints in the current stratum as will allow the installation of preferred constraints in the *next* stratum. To take the markedness >> IO-faith bias: if there is more than one set of minimal IO faithfulness constraints that can be installed in stratum n that will allow some markedness constraints to be installed in stratum n+1, then the learner chooses the set that allows the *most* markedness constraints to be installed. In their BCD version, this idea is enforced only with respect to markedness >> IO-faith, but it can be generalized here to help enforce the OO-faith >> markedness bias as well.³⁴

With this final piece, we can now see the prose version of the BCD algorithm I will be assuming for the rest of the dissertation. It is given in 96) below: while this

³⁴ See Prince and Tesar 2004's definitions of 'Smallest Effective F-Sets and Richest Markedness Cascades' for the details.

version is given in my own wording, I wish to be explicit that the BCD method itself is in no way innovated beyond that of Prince and Tesar (2004) – my additions are just to add the OO-faith bias as step 1, and to include the specific >> general IO-faith bias in the way that I have.

96) My BCD algorithm

Step 1: Install all OO-Faith constraints that prefer no Ls

- a) if any can be installed, move onto step 3
- b) if none are left to installed, move onto step 2
- c) if some are left but none prefers no Ls,
 - i) Install the smallest set of W-preferring Markedness constraints that will allow the installation of the most OO-faith constraints in the next consecutive strata³⁵, and move onto step 3

Step 2: Install all Markedness constraints that prefer no Ls

- a) if any can be installed, move onto step 3
- b) if none are left to installed, move onto step 4
- c) if some are left but none prefers no Ls,
 - i) Find the set of context arguments of all W-preferring IO-Faith constraints. If there is only W-preferer, install it and move onto step 3, otherwise
 - ii) Determine all the specific-to-general relations among their contexts, using the Context Arguments and Contingent Specificity tests, and find the resulting set of W-preferring IO-faith constraints with the most specific context arguments, and then
 - iii) Install the smallest set of these W-preferring IO-faith constraints that will allow the installation of the most Markedness constraints in the next consecutive strata and move onto step 3

³⁵ Again, see the definition of ‘Richest Markedness Cascade’ in Prince and Tesar (2004) to understand what “the next consecutive strata” means.

Step 3: Remove all resolved errors from the Support, and return to Step 1 to build the next stratum

Step 4: Install all remaining IO-faith constraints in the bottom stratum, and END.

Having adopted this algorithm, the reader whose primary interest is in natural L1 learning data can now safely skip ahead to the summary of this chapter in section 8. For others, however, the two sections below provide some discussion of the ways in which specific to general relations can be calculated, and their consequences for BCD’s ranking decisions.

7.2 Ways of calculating Spec-F >> Gen-F relations: the Azba case study

In this section, I illustrate the way in which the set of faithfulness constraints the learner chooses to calculate IO-faith specificity will affect the learning results. The case under discussion here is the Azba language, invented by Prince and Tesar and also discussed by Hayes. (It is a hypothetical example, but these authors note that it resembles both Attic Greek and Russian.) The example is one in which the learner is trying to learn the distribution of voiced fricatives, in a language where fricative voicing only appears in coda position through regressive assimilation from onset stops – e.g. [az.ba]. In all other contexts, the language only allows voiceless fricatives (e.g., [sa] and [as]), while stops can be either voiced or voiceless. As a result, the Azba lexicon looks like this:

- 97) *the schematic Azba lexicon*
- a) stops: [ba] [pa] [ab] [ap]
 - b) fricatives: [sa] [as]
 - c) clusters: [azba]
[aspa]

The concern is what faithfulness constraint will be used to protect the fricative's voicing in [az.ba].

In the illustration that Prince and Tesar originally provided and which I will follow, there are three Markedness constraints: two that penalize voiced obstruents (*b and *z), and one that requires voicing agreement between obstruent clusters (Agree(voice)). There are also four faithfulness constraints: defined just as Ident(b), Ident(z), and their onset-only versions, which we can construe as constraints protecting voicing.³⁶ With these constraints and this lexicon the learner's errors will all result from devoicing, in the three contexts that voicing appears: onset stops, coda stops, and coda fricatives. Thus, the Azba learner's set of ERC rows looks like this:

98) *The Azba learner's Support*

	Agree (voice)	*b	*z	Ident(b) -Onset	Ident(z) -Onset	Ident(b)	Ident(z)
(i) ab ~ ap	e	L	e	e	e	W	e
(ii) ba ~ pa	e	L	e	W	e	W	e
(iii) azba ~ aspa	e	L	L	W	e	W	W

From this Support set, the BCD learner will first install Agree(voice) in the top stratum. But since Agree(voice) assigns no Ws it resolves no errors, and both remaining markedness constraints *b and *z both assign Ls. So what IO-faithfulness constraint(s) should the learner choose?

³⁶ This would mean, in the terms of section 6, that these constraints are really Ident(voice)-fricative, Ident(voice)-stop, Ident(voice)-Fricative-Onset and Ident(voice)-Stop-Onset: these are the definitions used by Hayes. Since there is no specificity relation between the context arguments 'stop' and 'fricative', I will use the simpler constraint labels from Prince and Tesar. However, this point raises the issue of whether it is ever necessary to compute specificity using *both* prosodic and subsegmental contexts, and if so how these would be interpreted by the ranking algorithm (in particular, which would take precedence.) See § 7.4.

In the Azba language, the right constraint to install is Ident(b)-Onset, because it is the combination of onset stop voicing and undominated Agree(voice) that causes voiced codas. The issue is therefore whether the specificity bias alone will ensure that the learner chooses Ident(b)-Onset.

7.2.1 **Using a context-based F-specificity bias**

In the BCD algorithm that I defined in 96), the learner will indeed choose the right constraint. This is partly because my learner computes specific to general relations via tests across *contexts*, not constraints. It is not necessary for a more specific version of a particular constraint (like Ident(z)-onset) to assign Ws in order to notice that Ident(z) has a very general context. This is illustrated below: after having assigned Agree(voice) to its first stratum, my learner will use its third step to install a constraint in the second stratum and come up with the right choice:

99) *BCD learning of Azba with a specific-F contexts bias*
(Since there are no OO-faith constraints under discussion here, I skip Step 1)

To build stratum 1:

Step 2: Install all Markedness constraints that prefer no Ls
Resulting stratum: **Agree(voice)**

Step 3: Errors removed from the Support: none.
Error remaining: ba ~ pa, ab ~ ap, azba ~ aspa

To build stratum 2:

Step 2: Install all Markedness constraints that prefer no Ls
c) since some are left but none prefers no Ls:

i) Find the set of context arguments of all W-preferring IO-Faith constraints.

W-preferring constraints: {Ident(b)-Onset, Ident(b), Ident(z)}
Context arguments: {Onset, Seg}

ii) Determine all the specific-to-general relations among their contexts, using the Context Arguments and Contingent Specificity tests, and find the resulting set of W-preferring IO-faith constraints with the most specific context arguments.

Resulting relations: **Seg** is less specific than **Onset**

iii) Install the smallest set of these W-preferring IO-faith constraints that allows the installation of the most Markedness constraints in the next consecutive strata and move onto step 3

Resulting stratum: **Ident(b)-Onset**

Step 3: *Errors removed from the Support:* ba-pa, azba ~ aspa
Error remaining: ab ~ ap

Now everything is smooth sailing. The only remaining error is 98j), whose ERC row contains only one W-preferring IO-faith constraint, Ident(b). And after installing Ident(b) in the third stratum (by Step 3i), the learner will have resolved all the errors in the Support, so they are free to install all remaining markedness constraints in the fourth stratum (*b) and (*z), and then all remaining IO-faith constraints in the final stratum (Ident(z) and Ident(z)-Onset).

I spell out these remaining steps of the algorithm below just for completeness. But the crucial point is that this learner has installed all faithfulness constraints related to voiced fricatives (z) below the markedness constraint *z, and so chosen the most restrictive grammar consistent with the data:

100) *BCD learning of Azba with a specific-F contexts bias, part two*
(continuing to skip Step 1, as in part one)

To build stratum 3:

Step 2: **Install all Markedness constraints that prefer no Ls**
c) since some are left but none prefers no Ls:

i) Find the set of context arguments of all W-preferring IO-Faith constraints. If there is only one W-preferrer install it.
Resulting stratum: **Ident(b)**

Step 3: *Errors removed from the Support:* ab ~ ap
Error remaining: none.

To build stratum 4:

Step 2: **Install all Markedness constraints that prefer no Ls**
Resulting stratum: ***b, *z**

To build stratum 5:

Step 2: **Install all Markedness constraints that prefer no Ls**
b) since none are left to installed, move onto step 4:

Step 4: **Install all remaining IO-faith constraints in the bottom stratum, and END**
Resulting stratum: **Ident(z)-Onset, Ident(z)**

101) **The final Azba ranking:**
Agree(voice) >> Ident(b)-Onset >> Ident(b) >> *b, *z >> Ident(z)-Onset, Ident(z)

7.2.2 Using a constraint-based F-stringency bias: Hayes' simulation

Appendix A of Hayes (2004) discusses the Azba example, from both the perspective of his own Low-Faithfulness Constraint Demotion algorithm as well as a version of BCD that includes a version of the specific >> general faithfulness bias. He concludes there that if the BCD is equipped with this bias, the Azba-learning child will indeed correctly choose to install Ident(b)-Onset. His demonstration of this, using both algorithms, appears on his website.³⁷

³⁷ <http://www.linguistics.ucla.edu/people/hayes/Acquisition/AzbaSpecificityBCD.htm>

Hayes' version of the specific faithfulness bias is like the one I have proposed in that he determines the specificity of contexts by looking at winners, but the way the bias uses that specificity information is somewhat different. Once his learner determines that no more markedness constraints can be installed, it determines the *stringency relations* of the faithfulness constraints themselves – and not just of all W-preferring constraints, but of *all unranked IO-faithfulness constraints*. It then and rules out all the more general ones, before considering which of the remaining constraints are W-preferring.

In the Azba case, this means that both the general Ident constraints are ruled out because their Ident-Onset constraints are less stringent. Then, choosing a *W-preferring* faithfulness constraint necessarily picks the Ident(b)-Onset constraint, because Ident(z)-Onset does not prefer any winners (it only assigns es.)

To see that this works, I apply this approach below to just build the first two strata of the Azba grammar:

102) *BCD learning of Azba with an all-F stringent constraint bias* (adapted from: <http://www.linguistics.ucla.edu/people/hayes/Acquisition/AzbaSpecificityBCD.htm>)

a) *The Support, repeated*

	Agree (voice)	*b	*z	Ident(b) -Onset	Ident(z) -Onset	Ident(b)	Ident(z)
(i) ab ~ ap	e	L	e	e	e	W	e
(ii) ba ~ pa	e	L	e	W	e	W	e
(iii) azba ~ aspa	e	L	L	W	e	W	W

To build stratum 1

Step 1: Install all Markedness constraints that prefer no Ls
 Resulting stratum 1: **Agree(voice)**

To build stratum 2

Step 1: Install all Markedness constraints that prefer no Ls
 Resulting stratum 2: -- empty --

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

Step 2: Rule out all unranked Faithfulness constraints that are more stringent than any other.
 Constraints ruled out: Id(b), Id(z)
 Remaining constraints: Id(b)-Onset, Id(z)-Onset

Install a remaining Faithfulness constraint that prefers a W.
 Resulting stratum 2: **Id(b)-Onset**

103) **The first two strata:** Agree(voice) >> Id(b)-Onset

It is, of course, something of a lucky accident that there was only one remaining faithfulness constraint that assigned a W at stratum 2 – or rather, a function of the small constraint set used to illustrate this case. This just means we did not need to rely on the other principles for choosing among IO-faithfulness constraints in the later parts of Step 2 – or indeed the principles for choosing between IO-faith constraints used by Hayes 2004 to similar ends.

The important point about the Hayesian specific-to-general faithfulness bias is that it requires calculating the stringency relations of every unranked faithfulness constraint in CON. In the Azba case: the reason the learner knew not to install Ident(z) is because it knew Ident(z)-Onset was more specific – *even though Ident(z)-onset assigned no Ws*, and therefore could not possibly be useful in resolving errors. In the approach to calculating context specificity that I presented in section 6, this would mean recalculating a Context Table across ALL contexts, at least once per re-ranking (or at least across all contexts whose relationships that could be contingent.)

7.2.3 Prince and Tesar (2004) on the Azba language

A third approach would be to calculate the stringency relations between W-preferring constraints. This is the specific-over-general faithfulness bias that Prince and

Tesar (2004) consider, and which they point out will be unable to find the right constraint to install in the Azba case. As just emphasized above: looking just at the W-preferring constraints, the learner will not be required by an F-stringency bias to rule out Ident(z), because the specific Ident(z)-Onset constraint doesn't prefer any winners:

104) *BCD learning of Azba with W-preferring stringent constraint bias*

a) *The Support, repeated*

	Agree (voice)	*b	*z	Ident(b) -Onset	Ident(z) -Onset	Ident(b)	Ident(z)
(i) ab ~ ap	e	L	e	e	e	W	e
(ii) ba ~ pa	e	L	e	W	e	W	e
(iii) azba ~ aspa	e	L	L	W	e	W	W

To build stratum 1

Step 1: Install all Markedness constraints that prefer no Ls
Resulting stratum 1: **Agree(voice)**

To build stratum 2

Step 1: Install all Markedness constraints that prefer no Ls
Resulting stratum 2: -- empty --

Step 2: Find the set of W-preferring Faithfulness constraints

Constraints ruled out: Id(z)
Remaining constraints: Id(b)-Onset, Id(b), Id(z)-Onset

Step 3: Rule out all unranked W-preferring Faithfulness constraints that are more stringent than any other.

Constraints ruled out: Id(b)
Remaining constraints: **Id(b)-Onset, Id(z)**

7.2.4 Summarizing the Azba results

The Azba discussion above suggests that there are two good ways of calculating specific-to-general relations among IO-faith constraints: either we calculate across context arguments for a small(er) set of constraints as I have done, or we calculate across

a large(r) set of constraints as Hayes does. As I mentioned above, the approach that I use does have the potential benefit of requiring fewer calculations among contexts with potential contingent subset relations. But the real choice between these options (and any others) will have to be made by using them in learning algorithms to handle a wide variety of data.

7.3 Returning to Anti-Paninian rankings and phonotactic learning

The Azba example has brought out one further point about the workings of the specific-over-general faithfulness bias adopted here. What we've seen is that even when a general faithfulness constraint must be ranked above a markedness constraint, my learner will choose first to install the specific one, and then the general one. We can see this by focusing our attention on just the stop voicing constraints in the Azba example, in the Support and the final ranking:

105) *The Support, repeated in part*

	*b	Ident(b) -Onset	Ident(b)
(i) ab ~ ap	L	e	W
(ii) ba ~ pa	L	W	W
(iii) azba ~ aspa	L	W	W

106) *The final ranking learned, repeated in part from 101:*

Ident(b)-Onset >> Ident(b) >> *b

It is clear from the first error 105i) that general Ident(b) will have to be ranked above *b in the final ranking – and installing this faithfulness constraint would be

sufficient. In other words, our learner would get the right surface pattern if he had instead used a ranking algorithm that chose this ranking:

107) *A different ranking for Azba, that is equally surface-restrictive:*
 Ident(b) >> *b >> Ident(b)-Onset

Is there any problem with choosing 106) instead of 107) from this Support? One point already addressed in section 4.1.1 is that the ranking in 105) is sort of Anti-Paninian (Prince 1997 *et sqst*), in that it ranks a general faith constraint above a specific one with some constraint intervening.³⁸ So to rephrase the question: is there any problem with a learner who never adopts Anti-Paninian rankings from this Support?

Luckily, no. The crucial thing about AP rankings among faithfulness constraints is that they can *only* be necessary to analyze alternations – and alternations will provide a Support different than the one in 105). Here I will demonstrate this using the Swedish voicing case from Lombardi (1999)³⁹ as an example. I will not provide a full account of how the AP ranking should be learned, but merely suggest the aspects of the Support that would be relevant to its discovery.

From the perspective of phonotactic learning, the facts of Swedish voicing are just like Azba: coda-onset clusters can either be all voiced or voiceless. Using the by-now familiar constraints, the error in 108) shows only that *some* Ident[voice] constraint must out-rank *voice – so my BCD algorithm will choose the ranking in 109):

³⁸ Though it is not crucial that the general ranks above the specific – thus, it might be better to refer to these rankings as ‘incidentally Anti-Paninian’.
³⁹ This discussion owes much to Alan Prince’s LSA 2005 summer institute course notes.

108) *The Support for Swedish voicing in phonotactic learning*
 (using the Lombardi, 1999 analysis)

	Agree (voice)	*voice	Ident(vce) -Onset	Ident(vce)
(i) azba ~ aspa	e	L	W	W

109) *Resulting ranking*
 Agree[voice] >> Ident(vce)-Onset >> *voice >> Ident(vce)

However, this ranking turns out not to be the right one for Swedish. What alternations from Swedish demonstrate is that obstruent clusters can only be voiced on the surface if *both* of their input members were voiced, as in 110d). If either input obstruent was voiceless, the cluster will also be voiceless (110a-c):

110) *The truth about Swedish – from alternations*
 (a) /apta/ → [apta] (d) /abda/ → [abda]
 (b) /apda/ → [apta]
 (c) /abta/ → [apta]

The ranking that the phonotactic learner in 109) adopted, however, cannot explain the mapping in (110b): the fact that /apda/ devoices to [apta], even though its input onset member was voiced. Thus, assuming this ranking will lead to the following error:

111) *The error made by the phonotactic learning grammar:*

/apda/	Agree (voice)	Ident(vce) -Onset	*voice	Ident(vce)
⊖ apta		*!		*
⊖ abda			**	*

112) *The Support for Swedish voicing, including alternations*

<i>input</i>	<i>winner ~ loser</i>	Agree (voice)	*voice	Ident(vce) -Onset	Ident(vce)
(i) /azba/	azba ~ aspa	e	L	W	W
(ii) /apda/	apta ~ abda	e	W	L	W

At this point, our learner has evidence that Ident(vce)-Onset chooses a *loser*. This is something novel about learning from alternations – recall that in phonotactic learning, IO-faithfulness could only assign es and Ws. In the face of L-preferring faithfulness constraints, the F-specificity bias that I’ve suggested could be amended. For example, Step 3 could install faithfulness constraints that prefer *some winners and no losers* – this would correctly choose the general Ident(voice) constraint in 112).

Whatever the right strategy to deal with alternations, my present point is only that general >> specific rankings among faithfulness constraints cannot be necessary on the basis of phonotactics alone. Since the learner must already be assuming unfaithful I-O mappings before they’re crucial, this investigation will fall into the broad category of issues for restrictive learning in the face of alternations.

7.4 Summary and outstanding issues

In this section I have implemented my bias for the most specific IO-faithfulness constraints possible, using the tools I built in section 6 for calculating both universal and contingent specific-to-general relations between the *contexts* of faithfulness constraints. I have put this bias in the context of my version of BCD, which I will use in the next chapters. I have also provided some discussion of how this bias compares to the ones considered by Hayes (2004) and Prince and Tesar (2004), and raised the issue of how this bias will carry over to the learning of alternations (§7.3.)

There are of course many outstanding issues. One central issue is how the IO-faith specificity bias interacts with other principles that enforce restrictiveness in the ranking of IO-faith constraints – that is, whether the details of Step 2c) of my algorithm are the best ones. But remaining questions should also be asked about the workings of the bias itself. One already alluded to in footnote 33 is the potential interaction between degrees of specificity in prosodic and subsegmental contexts. That is: what should the BCD learner do when choosing between the following two constraints?

113) *Two potential faithfulness constraints with conflicting specificities*

- a) Ident[voice]-Stop
- b) Ident[voice]-Obstruent-Onset

The problem is that stops are more specific than obstruents, but onsets are more specific than segments. Thus, at the subsegmental level 113a)’s context is more specific, but at the prosodic level 113b)’s context is more specific. In such a case – which constraint should the learner install? A second issue is the possibility that multiple faithfulness constraints might need to be considered to determine the specific-to-general relations between context arguments.

In part, the answer to these questions will come from a better understanding of the correct theory of faithfulness. (In the first instance: if there is no constraint like Ident[voice]-Obstruent-Onset, that particular ranking indeterminacy will never arise.)

8. Chapter 2 Summary, in preparation for Chapter 3

This chapter has presented a view of error-driven phonotactic learning in Optimality Theory, using one of a class of Biased Constraint Demotion algorithms. I

have drawn together arguments from the literature to support three ranking biases for choosing the most restrictive grammar consistent with learning data. Together, these biases aim to rank OO-faith above Markedness, and Markedness above IO-Faith, and to install only the most specific of IO-Faithfulness constraints rather than any more general ones. I have focused in particular on this third bias, and argued that the difficulties in determining the most specific faithfulness constraints should be handled by calculating specificity across the *contexts* of faithfulness and examining the relations between those contexts in the language's observed output forms (winners.) And in discussing the benefits of BCD learning, I have continually stressed the use of stored errors – the Support – to ensure that restrictive grammars can be learned at every stage of acquisition, and so that early errors about the learning data do not persist in later rankings.

Most of the rest of this dissertation is concerned with how this BCD-style learner can be used to describe and predict aspects of natural language acquisition. Before embarking on this project, however, I note that Prince and Tesar are very explicit about the limited connection between their work on restrictiveness and the analysis of child data:

114) “It is important, however, to keep the subset issue notionally distinct from issues in the analysis of early acquisition patterns. The proposals we shall entertain are not intended to provide a direct account for child language data, although we expect that they ought to bear on the problem in various ways.” (Prince and Tesar 2004: 250)

With this caveat in mind – what aspects of child language data *could* this BCD learner speak to?

One prediction is that the biases of BCD characterize what is usually called the ‘initial state’. In other words, they provide the ranking of constraints before any errors

have been made and ERC rows added to the Support.⁴⁰ Thus, they also provide the grammar which begins the process of creating errors. For the BCD learner I gave in section 7.1, this initial ranking will look like 115):

115) *The initial ranking, chosen by my BCD algorithm in the absence of ERC rows*
OO-Faith >> Markedness >> IO-Faith

Note that the IO-faith specificity bias can't have any initial ranking effect; this is because IO-faithfulness constraints are only spread out in the ranking if errors demand it, and until then are dumped in one stratum at the bottom of the hierarchy by Step 4.

What does this ranking tell us about early grammars? The high Markedness bias predicts that early grammars will be unmarked compared to the target. The high OO-faith bias makes predictions only once morphology has been learned; I will return to some support for this prediction in chapter 4 with data from Kazazis (1969), Bernhardt and Stemberger (1998) and Smith (1973). And while the IO-faith specificity bias isn't relevant here, the existence of both specific and general faithfulness constraints at the bottom of our hierarchy predicts repairs in unprivileged contexts.

But what about stages of BCD learning once errors *have* been made? This is the subject of chapter 2.

⁴⁰ Shelley Velleman (p.c.) points out that this pure initial state is clearly not the state of the grammar at the advent of phonological production or word learning. By the time they begin to talk, or even to produce canonical babble, children have learned many language-specific aspects of their phonological grammar. I remain agnostic as to the point at which this knowledge is rightly represented using an OT grammar of the sort being assumed here, but it will at least be before meaningful speech production begins.

CHAPTER III
ERROR-SELECTIVE LEARNING

1. Introduction

Chapter 2 presented the case for OT learning via a Biased Constraint Demotion-style algorithm. While I have proposed some augmentation of Prince and Tesar (2004)'s original proposals with additional biases and associated calculations, the approach is still very much committed to Prince and Smolensky's Cancellation/Domination Lemma (chapter 2 ex. 3) as the method of learning. In other words, getting from the current grammar to the next always involves reasoning from a set of ERC rows to the rankings that will choose winners over losers.

Biased Constraint Demotion is a mechanism for learning *everything* necessary to find the right rankings; as we will now see, this means it is not designed in any way to learn gradually or imperfectly. As we just saw at the end of chapter 2, gradual, realistic human learning is *not* Prince and Tesar's goal – they carefully point out that their aim is a formal OT ranking algorithm that behaves in accordance with the subset principle, and not one that acts like a human child.

However: the issues of restrictiveness discussed in chapter 1 are ones that should not be ignored in the investigation of early child phonologies (recall the discussion of French stress from the beginning of chapter 2 §3.) To reach their target phonological grammar, real-life language learners must find a way to replicate the observed outputs that they hear while still being restrictive. To the extent that BCD is the best way we

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

know of keeping an OT learner restrictive, it therefore seems worthwhile to examine how a BCD learner could also be as gradual as a human child.

The goal of this chapter is therefore to propose a novel way of combining BCD with something that derives stages of acquisition – a way that relies on the insights of constraint demotion and ranking biases to derive some broadly-attested developmental stages in L1 learning. The proposal is called Error-Selective Learning: a method to make the learner choosy about *which* errors it adds to the Support, and therefore uses to re-rank via BCD.

This introductory section sketches Biased Constraint Demotion's over-quick learning, foreshadows the Error-Selective proposal, and provides two key examples of the attested intermediate acquisition stages that Error-Selective learning can derive. At the end of this section I provide a roadmap to the entire chapter, so that readers more interested in either data or theory can decide where to focus and where to skim.

1.1 The approach to reconciling BCD and gradual learning

The pure BCD learner performs its total re-ranking every time an error is added to the Support, which is every time an error is made. Since BCD is so efficient, it will learn everything there is to learn from each error as it is made, and ensure such an error is never made again.

Thus, one major way in which BCD is not a model of real-time language learning is that it fails to go through any intermediate stages: that is, it cannot learn *partially* from any of its data. For example, a learner whose current grammar does not permit any codas

to surface (described by the ranking fragment in (1)a below), only needs to make an error on a single word with a complex coda to acquire the fully-correct ranking in 1b):

- 1) *Starting state:* NoCoda, *ComplexCoda >> Max
Target state: Max >> NoCoda, *ComplexCoda

A representative flow chart of this process, using the English word ‘toast’, is sketched below:

2) *How the BCD learner gets from an error to a new grammar*

a) *At an Early Stage: this error is made*

/tost/	NoCoda	*CompCoda	Max
tost	*!	*	
tos	*!		*
to			**

adding error
to the Support

b) *... the error is put into the Learning Support Table...*

Input	Winner ~ Loser	NoCoda	*CompCoda	Max
/tost/	tost ~ to	L	L	W
/piz/	piz ~ pi	L	e	W

learning:
re-ranking
via BCD

c) *... and the next stage is the Final Stage, with no more coda errors*

/tost/	Max	NoCoda	*CompCoda
tost		*!	*
tos	*!	*!	
to	*!*		

There is no sense in which a BCD learner could choose to install Max above only NoCoda and not *ComplexCoda on the basis of the error in (2a). However, this is

precisely the kind of intermediate stage that learners of languages with complex codas (like English) do go through, often for several months of development (see §1.1.1).

The illustration in (2) shows that to avoid learning complex codas immediately from just one word like ‘toast’, our learner must either not learn using BCD, or it must not learn from errors like (2a). Since the previous chapter was dedicated to the claim that BCD is a successful way to ensure that learners reach the right end state grammar, the proposal I will make in this chapter takes the latter approach and provides a more gradual way in which errors enter the Support and so trigger re-ranking.

Briefly, the idea is this. In Error-Selective Learning, errors are not immediately added to the Support when they’re made, but rather stored temporarily until a sufficient number of ERC rows have demonstrated one particular problem with the current ranking. Only at that point does the learner choose to update the Support – and not with all errors made, but with just an error that will cause a minimal change to the grammar.

The criteria that choose the right error to add to the Support are designed with two particular kinds of intermediate stage in mind. In the following two sections (§1.2 and §1.3), I provide a representative example of each from the literature; many more examples of each appear in section 2 below for the data-interested reader. I acknowledge in advance that the two types of intermediate stages I discuss below do not exhaustively describe every such attested stage in the acquisition literature – and in fact, that there are patterns found commonly in development that will *not* be captured by the Error-Selective idea.¹ I focus here on these two stages in an attempt to make some initial progress, but the

¹ One aspect of phonological development I will have nothing to say about here is the role of child-specific templates in e.g. Priestly (1977), Vihman and Croft (XX), Vihman (XX) and references therein. These are templates that a particular child adopts at an intermediate stage which overwrites segmental material predictably across a range of lexical items, but which unlike most templatic effects in adult and other child

ultimate success of the Error-Selective model will have to be judged with respect to a much broader range of data.

1.2 The Specific Markedness stage: English coda clusters

The data below from Trevor (Compton and Streeter 1977 – see §2.1.1) demonstrate the claim made above that English-learning children often pass through three stages of coda acquisition. At the first stage no codas are produced, at the intermediate stage only singleton codas are produced,² and at the final stage complex codas now also appear.

3) Trevor's three stages of coda acquisition

a) All codas deleted (up to 1;4.2)

singleton codas			complex codas		
Target	Child	Age	Target	Child	Age
'duck'	[dʌ]	0;10.17	'plant'	[te]	1;3.11
'cup'	[kʌ]	1;1.0	'orange'	[oŋ]	1;4.2
'puppet'	[pʌpə]	1;3.25			

b) Intermediate stage: singleton codas only (1;5-1;7.26)

singleton codas			complex codas		
Target	Child	Age	Target	Child	Age
'walk'	[wɔk]	1;6.8	'box'	[gʌk]	1;7.11
'hat'	[hæt]	1;6.8	'toast'	[to:s]	1;7.20
'melon'	[mɛ:mm]	1;7.26	'milk'	[mʌ:k] ³	1;7.26

phonology do not seem to be motivated by markedness pressures: c.f. McCarthy and Prince (1993); Gnanadesikan (2004). There is also the tricky issue of apparent phonological regressions, which are at least partially addressed in §6.

² This is a simplification of Trevor's singleton coda development. I am abstracting away here from the fact that he actually appears to learn stressed codas before unstressed ones. See section 2.3.1 for more.

³ This lengthening of the vowel does not appear to be a consistent mapping for dark [ɪ]; many other words at this stage are transcribed with vowel lengthening that does not correspond to any missing input segment, and other missing [ɪ]s do not trigger vowel lengthening.

c) All codas retained (1;9 onwards)

singleton codas			complex codas		
Target	Child	Age	Target	Child	Age
'room'	[wu:m]	1;9.2	'plant'	[pænt]	1;9.2
'egg'	[eg]	1;9.28	'stairs'	[fɪtəz]	1;9.2
'outside'	[sai:d]	1;9.28	'toast'	[to:st]	1;9.29

The intermediate stage in 3b) is one which requires a ranking like in 4) below:

- 4) *A Specific Markedness stage*
*ComplexCoda >> Max >> NoCoda

As the simple tableaux below illustrate, this ranking protects singleton codas, but still reduces complex coda clusters:

5)a) Max >> NoCoda protects singleton codas in 'walk'

/wɔk/	Max	NoCoda
☞ [wɔk]		*
[wɔ]	*!	

5)b) NoComplex >> Max reduces coda clusters in 'toast'

/to:st/	NoComplexCoda	Max	NoCoda
[to:st]	*!		*
☞ [to:s]		*	*
[to]		*!*	

1.3 The Specific Faithfulness stage: French onset clusters

The second kind of intermediate stage comes from Rose (2000), who documented stages in the acquisition of Québécois French by two children, Clara and Théo. He presents evidence of a stage at which complex onsets are preserved faithfully in stressed syllables, but the same clusters are reduced to singleton in unstressed syllables:

6) *Clara's three stages of onset acquisition* (see Rose 2000:130-133)

a) All onsets reduced (1;0.28-1;09.01)

stressed syllable			unstressed syllable		
Target	Child	Gloss	Target	Child	Gloss
/kʁa.'kʁa/	[ka.'kæ]	'Cracra' (name)	/bʁi.'ze/	[bœ.'çi:]	'broken'
/plœʁi/	[pœ:]	'(s/he) cries'	/apʁi.'ko/	[pʁæ.'ko]	'apricot'
/flœʁ/	[βœ:]	'flower'			

b) Intermediate stage: stressed onsets retained (1;09.29-2;03.05)

stressed syllable			unstressed syllable		
Target	Child	Gloss	Target	Child	Gloss
/bi.'bʁɔ̃/	[pa.'pʁɔ]	'baby bottle'	/fʁi.'go/	[bu.'ko]	'fridge'
/gʁi/	[kʁi]	(he/she)'slides'	/bʁy.'le/	[bʁi.'le]	'burned'
/si.'tʁuj/	[θə.'tʁu:j]	'pumpkin'	/gʁi.'sad/	[ka.'sæd]	'slide'
/plœʁi/	[pʁœʁ]	'(he/she)cries'	/tʁu.'ve/	[tu.'ve]	'found'

c) All onsets retained (2;03.15 onwards)

stressed syllable			unstressed syllable		
Target	Child	Gloss	Target	Child	Gloss
/gʁo/	[gʁo]	'big'	/tʁu.'ve/	[tʁu.'ve]	'found'
			/plā.'fe/	[plā.'fe]	'floor'

The ranking this 6b) intermediate stage suggests:

7) *A Specific Faithfulness stage*
 Max-(σ') >> *ComplexOnset >> Max

In this ranking, markedness is sandwiched between two faithfulness constraints – the higher-ranked of which is a positional version of Max. I will return to the correct definition of this constraint in section 1.4, but its effect will be to prevent input segments from being deleted from stressed syllables. With this constraint, we get Clara's intermediate stage, as below:

8)a Max σ' >> *ComplexOnset protects clusters in stressed syllables

/gʁi/	Max-σ'	*ComplexOnset
kʁi		*
gʁi	*!	

8)b *ComplexOnset >> Max reduces clusters elsewhere

/gʁi.'sad/	Max-σ'	*ComplexOnset	Max
kʁa.'sæd		*!	
ka.'sæd			*

It should be noted that the specific faithfulness constraints that I focus on in this chapter are positional rather than featural – that is, the contexts in which they apply are prosodic categories higher than the segment (stressed syllables, initial syllables, and Roots). However, many OT analyses of intermediate stages in child development use Ident[feature] constraints in ways that, if translated into a *stringent* theory of featural faithfulness, would provide parallel rankings to the one in (7). (For examples, see Pater (1997) section 4.3 on child consonant harmony constraints intermingled with faithfulness and the grammar fragments in Pater and Barlow (2003) figures 2-4.)

1.4 Analytic assumptions about the intermediate stages

1.4.1 Stringency relations among markedness constraints

Chapter 2 already made it clear that the theory of faithfulness I adopt uses stringency relations to capture the positional and specific contexts of faithfulness constraints. In the coda vs. complex coda example from 1.2 above, I've similarly characterized the "flanking" markedness constraints as being in a stringency relation. The top Markedness constraint is *ComplexOnset, given as a separate constraint from *ComplexCoda – rather than using a single *Complex constraint which would not be in a

stringency relation with NoCoda (because CCVC syllables would violate the former and not the latter.).

In this particular case the split seems well-justified, if only because children acquire clusters edge by edge (see the data above, or the Dutch data presented in Levelt and van der Vijver, 2004). There might well be different (or even better) analyses of the data that I present in this chapter that does not involve stringency relations between its constraints – and certainly there are intermediate stages that do not include such stringency relations. In the end, stringency relations between markedness constraints will turn out to be a relevant but not necessary aspect of my Error-Selective method of deriving stages. However, they do provide a clear example of how the proposal works.

1.4.2 Positional faithfulness and input prosodic structure

Two kinds of positional faith constraints are used in this dissertation to capture intermediate stages of acquisition. The first is the more common positional Ident constraint used normally in analyses of adult grammars (see the body of references in chapter 1 §4.1). As we already saw in chapter 1, these constraints retain input properties in a particular output context, e.g.:

- 9) Ident[voice]-Onset: “Output segments syllabified as onsets must match their input correspondents for the feature [voice]”

The second kind of positional faithfulness constraint, which in fact play a role in the majority of this section’s analyses, are Max constraints with a positional flavour. These constraints prohibit deletion only in certain output contexts. To make coherent the

notion of “deleting from an output context”, these constraints must be defined across inputs and outputs which *both* contain prosodic contexts.⁴ Thus the Max-σ’ constraint used in section 1.3 will be defined as:

- 10) Max[Seg]-Onset: “Input segments syllabified as onsets must have output correspondents”

These constraints are therefore only violated when segments are already syllabified as onsets in the input and then deleted in the output.⁵

As a consequence of their definition, these positional Max constraints require the assumption that the learning child has prosodic structure in his or her *inputs* as well. In the French case of complex onsets above: Clara must first have gotten the French syllabification right, to have parsed word-medial obstruent-liquid clusters as complex onsets. But she must also have encoded that syllabification in the input itself, so that the constraint in (10) protects her input stressed syllables.

The problem with assuming input prosodification and positional Max constraints of this sort is that they will predict an unattested set of languages with contrastive syllabification – e.g. a language that contrasts [pa.ta] and [pat.a] to remain faithful to input syllable structure (for discussion of this point, see e.g. McCarthy 2006). This state of affairs presents a dilemma: on the one hand it appears that adult grammars do not

⁴In fact, positional Ident constraints may need to refer to input prosodification in order to account for a wider range of phenomena than those discussed in this chapter -- see especially Wilson (2000) – although see McCarthy (2006) for a solution to these problems for Ident that involves a different notion of GEN (OT with Candidate Chains.) Whether an OT-CC approach can also provide a solution to the problem with positional Max raised in the main text above is an interesting question for further consideration.

⁵Note that Beckman (1998) chapter 5 defines Positional Max rather differently; see also Alber (2001).

syllabify contrastively, while on the other hand child grammars DO require positional constraints that ban deletion (see the data §2.3 below).

To resolve this conflict, something learning-specific must be said about positional Max constraints.⁶ One possible approach would be to say that learners are prevented from positional deletion by *output-output* faithfulness relations of some sort – e.g. requiring segments in strong contexts of the *winner*s to be retained in the *loser*s.⁷ In this way, we might re-define the Max[Seg]-Onset constraint in 10) above as 11):

- 11) OO- Max[Seg]-Ons: “*Winner segments syllabified as onsets* must have output correspondents.”

Such positional Max constraints would avoid any typological problems among adult grammars, because they can only assess violations in a tableau when one candidate is designated the *winner*.⁸

To prevent distraction from the central points of the chapter, I will not adopt this OO-faith definition in 11) in the tableaux to come, and instead will add prosodic structure to the learner’s inputs when necessary. However determining the correct definition of positional Max constraints and their precise learning-specific properties are crucial issues for future research, especially because they raise important questions about the precise status of winners and losers and the relationships between developing and end-state grammars.

⁶ Cf. the discussion of this approach by Rose (2000); see also the child-specific syllabic structures assumed by Goad and Rose (2004).

⁷ Thanks to John McCarthy for initial discussion of OO-faith’s role in defining these constraints, and to Della Chambless and Joe Pater for later discussion.

⁸ Assessing this constraint is admittedly not a trivial matter – it would mean understanding ‘winner’ as something like a Fully-Faithful Candidate (see McCarthy 2006) that stands in correspondence with every other candidate. This is also reminiscent of the sympathetic candidate, in Sympathy Theory approach to opacity (McCarthy 1999)

1.5 Roadmap to the chapter

To provide some fodder for the proposal, section 2 below provides a small range of examples of the two kinds of intermediate stage discussed above, from both the existing literature and some of my own corpus work. Section 2.2 concentrates on a couple of Specific-M cases; section 2.3 spends some more time introducing a range of Specific-F effects. (The less data-minded reader can skip this section without theoretical repercussions.) Section 3 then introduces the Error-Selective Learning technique: spelling out the procedure in §3.1, discussing some of why and how it works in §3.2, and exemplifying its use in a case study of two children’s complex onset acquisition in §3.3.

Section 4 turns to a more thorough discussion of one aspect of Error-Selective Learning, namely the way its stages are connected to input frequencies. I discuss the frequency predictions of ESL (§4.1), and some evidence for the predictions from the literature (§4.2), including predictions that are not strictly tied to the Specific-M and Specific-F stages (§4.3). I also point out that ESL’s partial reliance on error frequency makes the BCD learner robust to noisy data. Section 5 makes some speculative proposals about how the Error-Selective CD learner could be extended to model variation between stages of acquisition, and section 6 concludes.

2. The data from intermediate stages

2.1 Introduction to the data

The data discussed in this chapter has been selected both to demonstrate the kinds of intermediate stages that Error-Selective learning can handle, but also to provide samples of representative intermediate stages in the literature to date. Much of the chapter focuses on the acquisition of syllable structure: codas and coda clusters, onset clusters

and other syllable margins, although examples are also drawn from the development of word shape (i.e. stages of syllable truncation). Later, in chapter 4, I will also turn briefly to some attested stages of morpho-phonological development (chapter 4 §7.2; see also this chapter §2.3.4).

I note here that the stage-by-stage characterization of the data in this section will abstract away from any and all variation within stages (although clearly the quantitative results I provide and cite from others will belie this idealization.) I leave the discussion of variation to section 5.

2.1.1 The Compton/Streeter database

Much of the data discussed in this chapter is taken from existing sources (see references with each example) but I also use data from my own work on the corpus of two children, Trevor and Julia. These data are taken from the Compton/Streeter/Pater database (Compton and Streeter 1977; Pater 1997; Pater and Werle 2001; Pater and Barlow 2003.) This database contains transcriptions by the children's mothers, who were speech pathologists and had received additional training in child transcription before beginning the data collection. Data on the amount and breadth of the data from the children is provided below in 12):

12) *Statistics about Trevor and Julia's corpora*

	first and last session	total no. of tokens ⁹
Trevor	0;8 – 3;1.8	12,177
Julia	1;2 – 3;1.3	5,772

⁹ Note that some 'tokens' are really utterances, so that one 'token' might include more than one word in the data reported below.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

Table 13) provides the number of different words that each child had used by the end of each month they were followed:

13) *Total lexical items*

up to...	Trevor	Julia
1;3	78	--
1;4	129	30
1;5	196	52
1;6	297	112
1;7	374	164
1;8	440	232
1;9	484	309
1;10		379

The data for these two children seems rather comprehensive – that is, it would appear to contain nearly all the words (types) Trevor and Julia uttered, up until near the end of their second year (1;9 or 1;10).

2.2 Intermediate stages that rely on specific markedness

2.2.1 More on complex codas in Germanic

It is well-established that children acquiring languages with complex coda structures (like English and Dutch) often go through an intermediate stage where singleton codas are preserved faithfully, while coda clusters are reduced to singletons.

14) <i>The initial stage:</i>	<i>The intermediate stage:</i>
/CVC/ → [CV]	/CVC/ → [CVC], *[CV]
/CVCC/ →	/CVCC/ → [CVC], *[CVCC]

This stage was exemplified with some of Trevor's data in section 1; in (15) below I provide additional examples from other children at this intermediate stage. The Dutch

data from Eva come from Fikkert (1994), Levelt (1994); G's English data come from Gnanadesikan (1995/2004), and P.J.'s English data are from Demuth and Fee (1995):

15) *The intermediate stage of coda acquisition*

	singleton codas retained			complex codas reduced		
	Target	Child	Gloss	Target	Child	Gloss
Dutch: Eva (1;4,12)	/te:n/	[ten]	'toe'	/e:n/	[ein]	'duck'
	/bed/	[deɪ]	'bed'	/sta:rt/	[taɪ]	'tail'
English: G (2;3-2;9)	/gre:p/	[gep]	'grape'	/drɪnk/	[bɪk]	'drink'
	/pi:z/	[piz]	'peas'	/frend/	[fen]	'friend'
	singleton codas retained (or deleted)			complex codas reduced (or deleted)		
English: PJ (1;11)	/wɔ:k/	[rɔ:]	'walk'	/tɔ:st/	[to:s]	'toast'
	/sup/	[sup]	'soup'	/bidz/	[bi:s]	'beads'
		[su:], [su]			[be:]	
	/dʒus/	[dʒu:s], [du:s]	'juice' ¹⁰			
[dʒu:]						

Again: this pattern is derived by the ranking in 16):

16) *ComplexCoda >> Max >> NoCoda

2.2.2 **Markedness of complex onsets, and sonority distance**

A different example of a stage that requires this kind of ranking comes from the development of onset clusters. Along the developmental path from singleton onsets to the full English set of complex onsets, Trevor and Julia both go through stages where stop-r

¹⁰ In Demuth (1996) the child pronunciation AND adult target are transcribed as [dz], rather than [dʒ]. I assume this is a typo, but in any event the quality of the child's *onset* is not important here.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

and stop-glide onsets consistently surface faithfully, but stop-l onsets are almost always reduced to singletons:

17) *The initial stage:* all /CVC/ → [CVC] *The intermediate stage:* /C_rVC/, /C_wVC/ → [CCVC], *[CVC] /C_lVC/ → [CVC], *[C_lVC]

As summarized in 18), Trevor's initial stage lasts until 2;2:

18) Trevor's initial stage: all onset clusters reduced (up to 2;2)

age	complex onset inputs			
	outputs raw #s		output percentages	
	[C]	[CC]	[C]	[CC]
up to 1;7	344	3	99.1	0.9
1;8-2;2	838	130	86.6	13.4

19) Trevor's representative words at stage one

stop-liquid clusters: reduced			stop-r (and stop-glide) ¹¹ clusters: retained		
Target	Child	Age	Target	Child	Age
'blocks'	[gak]	1;5.18	'cracker'	[kaka]	1;5.14
'clock'	[kak]	1;6.17	'train'	[ten]	1;5.14
'glasses'	[ˈgæ:fɪʃ]	1;8.12	'brush'	[baɪʃ]	1;6.25

After leaving the initial stage, Trevor then went through two months (2;3-2;4) at the intermediate stage discussed above. I have broken down Trevor's cluster treatment at this point into three categories: stop-r, stop-l and tr. I leave out tr clusters below, because Trevor's /tr/ cluster has something of an independent trajectory that seems mostly due to the [tʃr]-initial pronunciation of his own name that he adopts at this point.

¹¹ I have not included raw numbers from Trevor's input stop-w words at this stage because they are so few.

20) Trevor's intermediate stage: stop-r retained more than stop-l (2;3-2;4)¹²

age	output	raw #s		percentages	
		stop-l	stop-r	stop-l	stop-r
2;2	C...	30	34	100.0	82.9
	CC...		7	0.0	17.1
2;3	C...	23	12	69.7	31.6
	CC...	10	26	30.3	68.4
2;4	C...	28	18	62.2	41.9
	CC...	17	25	37.8	58.1

21) Trevor's representative words at the intermediate stage:

stop-liquid clusters: reduced			stop-r/stop-glide clusters: retained		
Target	Child	Age	Target	Child	Age
'glass'	[gʌs]	2;3.22	'grapes'	[grɛpts]	2;3.4
'play'	[peɪ]	2;3.30	'quite'	[kwait]	2;3.4
'cleaner'	[ki:nə]	2;4.13	'present'	[pwɛsɛn]	2;4.3

Julia (who talks much less than Trevor, but also begins to produce complex onsets much earlier) is at the initial stage of onset acquisition up until about the end of 1;9:

22) Julia's initial stage: up until 1;9.25

age	complex onset inputs			
	output raw #s		output percentages	
	[C]	[CC]	[C]	[CC]
up to 1;7	55	0	100.0	0.0
1;8-1;9	110	28	79.7	20.3

23) Representative words at Julia's stage one

stop-liquid clusters			stop-r and stop-w clusters		
Target	Child	Age	Target	Child	Age
'please'	[pis]	1;7.7	'cry'	[kai]	1;7.9
'blankie'	[bækɪ]	1;6.15	'drive'	[waɪv]	1;9.14
'clown'	[klaʊn]	1;8.25	'truck'	[fɹʌk]	1;9.25

¹²Two notes about this stage. First, it is sometimes Dep that gets violated in the case of stop-liquid clusters, rather than Max: 'problem' as [pwa:bələm], (2;3.7). Second, Trevor goes through a brief period at 2;3 where /p/ in particular is preserved as [pw], but then he reverts again to singleton [p].

Julia's intermediate stage lasts from about 1;10-2;1. At stage two I have broken down her clusters into stop-r, stop-l and br. This last cluster is separate because [br]'s acquisition is delayed compared to all other stop-r clusters (see the final two columns table of 24) below.)¹³ I have also included the numbers from 2;2, which show she's on her way to a stage where all these clusters surface:

24) Julia's intermediate stage: stop-r clusters retained while stop-l reduced (1;10-2;1)

age	output	raw #s		percentages		br	%
		stop-l	stop-r/w ¹⁴	stop-l	stop-r		
1;10	C...	21	9	95.5	26.5	4	
	CC...	1	25	4.5	73.5	0	
1;11	C...	22	1	100.0	2.4	4	
	CC...	0	40	0.0	97.6	2	
2;0	C...	23	3	100.0	7.9	5	
	CC...	0	35	0.0	92.1	0	
2;1	C...	17	2	89.5	5.4	4	
	CC...	2	35	10.5	94.6	0	
2;2 (beginning of next stage)	C...	0	5	0.0	10.4	0	
	CVCV	6	0	21.4	0.0	6	
	CC...	22	43	78.6	0	89.6	

25) Representative words from Julia's intermediate stage

stop-liquid clusters: reduced			stop-r/-glide clusters: retained		
Target	Child	Age	Target	Child	Age
'glasses'	[gʌθʌs]	1;10.10	'drink'	[grɪŋk]	1;10.5
'please'	[pis]	1;10.10	'queen'	[gwiŋ]	2;0.2
'clap'	[kæp]	1;11.15	'crown'	[kwaʊn]	2;0.2

¹³With respect to /br/: given the fact that she pronounces many, if not most, of these dependent onset /r/s as [w], one reasonable explanation is that her grammar rules out an onset cluster parse /br/ → [bw] via the OCP-labial constraint that independently rules out [bw] and [pw] in English.

¹⁴During this stage, Julia uses what is transcribed as [fw] for /sw/ and occasionally other clusters like /kw/. Noting that all these input clusters are voiceless, and include both labial and velar place, Joe Pater (p.c.) suggests Julia is actually producing a voiceless labio-velar consonant, having fused these clusters' features into a single segment.

This stage can be derived by ranking faithfulness (Max-Seg) between some two markedness constraints that affect complex onsets. Here I will suggest that the relevant constraints are on the relative sonority of the onset's segments. Below I spend some time on the theoretical details of these constraints because they will be used in the case study in section 3.3.

The constraints I will use here are an adaptation of Baertsch (2002)'s Split Margin Hierarchy constraints, which penalize onset clusters (among others) according to the sonority of first and second members.¹⁵ The sonority hierarchy I adopt is in 25) (see e.g. Blevins, 1995; Clements, 1990; Murray and Venneman, 1983; Parker, 2002; also Pater and Barlow, 2003):

- 26) *Relevant Sonority Hierarchy (from most to least sonorous)*
vowels > glides > r > l > nasals > fricatives > stops

Note that I have adopted a sonority hierarchy that distinguishes [r] as more sonorous than [l], precisely because it is this difference that matters here.¹⁶

Two well-known generalizations are (a) onset clusters rise in sonority, so that their first member is less sonorous than their second, and (b) the larger the rise in sonority, the better the cluster.¹⁷ To capture these generalizations, Baertsch (2002) expands on Prince and Smolensky (1993)'s Peak and Margin hierarchies and uses them similarly to build constraints which each penalize a sequence of sonority levels. Baertsch

¹⁵ See also Gouskova (2001) for a similar hierarchical OT approach to the markedness of cluster sonority, but for coda-onset sequences.

¹⁶ One argument for r's higher sonority than l comes from English rime structure, on the assumption that the more sonorous a segment the better a nucleus it makes. While some English speakers claim to have monosyllables in 'earl' and 'squirrel', none report the intuition of a monosyllabic [lr] rime sequence. See also Parker (2002).

¹⁷ at least to a point; see e.g. Clements (1990), in which the best sonority distance is calculated both between onset segments and between those segments and the following vowel nucleus.

then organizes her constraints into fixed rankings to reflect the second generalization above – the less sonority rises between the first and second members of an onset cluster, the more marked the cluster is, and the higher its constraints sit in the fixed ranking. This is illustrated in 27) below – I have adopted Gouskova (2004)'s adaptation of the constraints, though using the reduced sonority scale of 26) above. In these constraints, 'T' stands for any stop, 'S' for any fricative, 'N' for any nasal, 'L' and 'R' for themselves, and 'W' for any glide:

- 27) *Onset Sonority Distance Hierarchy (Baertsch, 1998, 2002; from Gouskova, 2004)*
*WT>>{*WS,*RT}>>...{*WW,*RR,*LL,*NN,*SS,*TT}>>...{*SW,*TR}>> *TW

Thus, the most marked onset cluster is one which maximally *falls* in sonority – $_{\sigma}[\underline{w}t\alpha]$, meaning in this case any glide followed by any stop – and the least marked is one which maximally *rises* – $_{\sigma}[\underline{twa}]$.

From the present perspective, I translate the constraints in (27) in the fixed hierarchy into a set of *stringent* constraints. These stringent constraints each ban a particular point on the onset sonority distance hierarchy, as well as anything above (i.e. more marked than) that point. This is shown in the prose definitions of these constraints below; note that not every point on the scale is illustrated on this scale (skipped constraints are indicated by the ellipses):

- 28) *Stringent Onset Sonority Distance Constraints (built from 27)*

- (a) *WT = "No glide-stop onsets"
(b) *WS, RT = "No glide-fricative or r-stop (or glide-stop) onsets"
(...)

- (c) *WW,RR,LL,NN, SS,TT
= “No onsets with a sonority plateau
(or with any sonority drop)”
(...)
- (d) *LW, NR, SL, TN = “No liquid-glide, nasal-liquid, fricative-liquid or stop-nasal
onsets (or anything with less or a sonority rise)”
- (e) *NW, SR, TL = “No stop-liquid, fricative-r, or nasal-glide onsets
(or anything with less of a sonority rise)”
- (f) *SW, TR, = “No fricative-glide or stop-r onsets
(or anything with less of a sonority rise)”
- (g) *TW = “No stop-glide onsets
(or anything with less of a sonority rise)”
i.e., “No complex onsets”

For the case we are analyzing here, the relevant two constraints are (28e) and (f); these are near the most stringent end of the constraint set, meaning they penalize all but the best clusters. Since the data we are focusing on here is just *stop-initial* clusters, we can reword the effects of these two constraints as in (29):

29) *Stop-initial clusters allowed by two onset sonority constraints*

- a) *NW, SR, TL obeyed by stop-r and stop-glide clusters
(abbreviated to *TL)
- b) *SW, TR, obeyed by stop-glide clusters
(abbreviated to *TR)

With Max-Seg between these constraints, we get the right effect:

30) *TL >> Max[Seg] >> *TR

To illustrate using Julia’s data from 25):

31) Max >> *TR protects the onset cluster in ‘drink’

/[driŋk]/	Max	*TL
^σ [gwiŋk]		*
[griŋk]	*!	

32) *TL >> Max reduces the cluster in ‘please’

/[plis]/	*TL	Max	*TR
[plis]	*!		*
^σ [pɪs]		*	

In this theory, the constraint *ComplexOnset can be dispensed with – or rather, it is simply equivalent to the most stringent onset sonority constraint, *TW. However, in all subsequent tableaux when onset sonority is not at issue, I will continue to use the simple constraint label *ComplexOnset.

2.3 Intermediates stages that rely on specific faithfulness

This section discusses developmental stages that are best (or only) analyzed with reference to positional faithfulness constraints – and in particular, positional Max.

2.3.1 More on faithfulness in stressed syllables

The first case, already discussed in section 1, comes from Rose (2000). He presents evidence of a stage at which complex onsets are preserved faithfully in stressed syllables, but the same clusters are reduced to singleton in unstressed syllables (see Rose 2000:130-133):

- 33) *the initial stage* /CV.'CCV/ → [...CVC...]
the intermediate stage /CV.'CCV/ → [CV.'CCV], *[CV.'CV]
/CCV.'CV / → [CV.'CV], *[CCV.'CV]

Note that at 1;9, all the “codas” that José is producing are *vowels*. It should be noted that Lléo does report explicitly that these vowels only ever appear as coda substitutes, and that he is not merely at a stage where vowel length/quality is uncontrolled. Nevertheless: while it may be that these epenthetic vowels are related to the input codas, they still do not indicate mastery of a stage where NoCoda has been demoted. Instead, this begins at 2;0 and increases considerably at 2;2:

40) Jose’s stage two – 2;0-22¹⁹

a) final codas

age	stressed σ			unstressed σ		
	targets	faithful	% faith	targets	faithful	% faith
2;0	31	4	13%	14	1	7%
2;2	30	6	20%	11	0	0%
total	145	18	12.4%	48	1	2.1%

b) medial codas

age	stressed σ			unstressed σ		
	targets	faithful	% faith	targets	faithful	% faith
2;0	74	35	47%	32	3	9%
2;2	344	69	20.1%	132	11	8.3%
total	83	12	14%	39	4	10%

Thus we have the ranking in 41); the tableaux in 42) below use data from Lléo

(2003) table 2 to demonstrate the ranking:

41) Max[Seg]-stressed- σ >> NoCoda >> Max[Seg]

¹⁹ As Lléo (2003) rightly notes, José’s production of codas in medial position is better than in final position – something I have nothing say about here.

42) NoCoda >> Max reduces the coda in *dos* ‘two’ (1;7.27)

/[dos]/	NoCoda	Max
[dos]		*
\varnothing [dœ:]	*!	

43) But Max- σ >> NoCoda retains the coda in *venga*, ‘come on’ (1;10.3)

/[benga]/	Max- σ	NoCoda	Max
\varnothing [benga]		*	
[bega]	*!		*

This stage also appears in Trevor’s data. During his first months of solid coda production, his outputs almost always retain codas only in stressed syllables. Given the difference in frequent word shapes of English vs. Spanish, I characterize his intermediate stage slightly differently than José’s, but the ranking remains the same:

44) *the initial stage*
 /CVC/ → [CV]
 /CVCVC/ → [CVCV]
the intermediate stage:
 /CVC/ → [CV], *[CV]
 /CVCVC/ → [CV], *[CV]²⁰
 /CVCVC/ → [CVCV], *[CVCVC]

45) Trevor’s stage one: no codas anywhere (up to 1;3)

stressed syllables			unstressed syllables		
Target	Child	Age	Target	Child	Age
‘duck’	[dʌ]	0;10.17	‘puppet’	[pʌpə]	1;3.25
‘cup’	[kʌ]	1;1.0	‘orange’	[oŋ]	1;4.2

46) Trevor’s intermediate stage: singleton codas only in stressed syllables (1;5-1;6)

stressed syllables			unstressed syllables		
Target	Child	Age	Target	Child	Age
‘all gone’	[gəˈɡɒn]	1;5.4	‘puppet’	[pʌpə]	1;5.5
‘bike’	[ɡaɪk]	1;5.30	‘blanket’	[kækt]	1;5.14
‘hat’	[hæt]	1;6.8	‘yogurt’	[ɡɒɡə]	1;5.30

²⁰ Truncation due to e.g. Trochee and Parse- σ

2.3.2 Faithfulness to stressed syllables

There are also several examples from the literature on syllable truncation where children resist the pressure to delete syllables from a privileged (stressed or initial) position.²¹

With respect to the stressed syllable position, Kehoe and Stoel-Gammon (1997) and Kehoe (2000) report on an elicitation study of English-speaking children at (2;4) and (2;10), designed in part to test for stress effects on syllable truncation. In their data, truncation patterns were almost exclusively restricted to unstressed syllables while stressed ones were retained. The most compelling evidence for Max-σ' in this data comes from a stage in 47) below, from Kehoe (2000)'s section 4.2.2:

47) *the English intermediate stage*
 /wSw/ → [(Sw)] /SwSw/ → [(S)(SW)]

48) *the data* (taken from Kehoe (2000) tables 4,7, 10)²²

	Subject	Target	Child	Target	Child
		/wSw/	[Sw]	/SwSw/	[SSw]
a)	22m3	banana	[næ̃nΛ], [næ̃ŋΛ]	àlligátor	[æ̃gè.Λ]
b)	22f1	banána	[næ̃nΛ], [næ̃nə]	àlligátor	[æ̃gæ̃də̃]
c)	27m6	banána	[báni]	àvocádo	[àkádo]
d)	28f2	banána	[bæ̃:ml]	àvocádo	[λkádo]

As Kehoe points out, output forms like [(à)(kádo)] for “avocado” include a marked initial degenerate foot (à). The explanation for why these children’s grammars preserve the first syllable of “avocado” but still truncate the first syllable of “banana”

²¹ Recall from section 1.4.2 that this analysis assumes the positions exist in the *input* as well.
²² Since this was a cross-sectional study, we don’t have the data to show any of the earlier truncation stages these children were at: e.g., one where outputs were always one foot (see §4.3.)

must therefore be a pressure to retain stressed syllables only, at the expense of marked foot shape. Thus, this stage provides evidence of another Specific-F ranking, namely:

49) Max[Seg]-σ' >> *DegenerateFoot²³ >> Max[Seg]²⁴

50) Markedness rules out a one-syllable initial foot in ‘banana’

/bənæ̃nΛ/	*DegenerateFoot	Max[Seg]
σ' (næ̃nΛ)		*
(bə̃)(næ̃nΛ)	*!	

51) Max(Seg)-σ' protects the initial syllable of ‘avocado’

/ävəkádo/	Max[Seg]-σ'	*DegenerateFoot	Max[Seg]
(kádo)			*
σ' (à)(kádo)		*!	

2.3.3 Faithfulness to initial syllables

One piece of evidence of an intermediate stage that relies on *initial* syllable faith comes from a later stage of syllable truncation in Greek (Revithiadou and Tzakosta, 2004); there is also some evidence for this stage in Spanish mentioned in Gennari and Demuth (1997). The former authors provide evidence from four different children acquiring Greek at a stage where words with more than 3 syllables get reduced to the stressed-foot, *plus the input’s initial syllable* – while other unstressed syllables are reduced:

²³ This choice of markedness constraint is not the only option; each option will probably require some other assumptions about footing at this stage. In this case, using *DegenerateFoot to rule out the initial syllable of ‘banana’ requires the assumption that Prosodic words must begin with a foot (i.e. that Align-Ft-L is undominated) – but this does indeed seem to be the case for many early stages of English prosodic acquisition: see e.g. Kehoe (2000), and also §4.3 of this chapter.
²⁴ John McCarthy suggests that this stage could also arise if the learner assumes that the initially-stressed syllables are long and therefore do not constitute degenerate feet.

- 52) *the Greek intermediate stage:*
 /wSw/ → [w(Sw)], *[(Sw)]
 /wwSw/ → [w(Sw)], *[ww(Sw)]

- 53) *the data (taken from Revithiadou and Tzakosta (2004, ex. 4))*²⁵

Subject	Target	Child Output	Gloss
B1 (2;09.25)	/ka_la.má.ci/	[ka:(má.ci)]	'straw-diminutive'
B1 (2;09.12)	/yu_ru.ná.ca/	[yu.(ná.ca)]	'pigs-diminutive'
D (2;04.05)	/me.li.ti.ni/	[me.(ti.ni)]	(name)
D (2;04.05)	/fo.to.yra.fi.es/	[fa.(fi.eθ)]	'photographs'

As with Kehoe's banana vs. avocado data, this truncation pattern also requires the use of a faithfulness constraint relative to initial syllables, because what *markedness* constraint could be prompting the retention of an initial unstressed syllable? Thus, sandwiching a markedness constraint against unfooted syllables, Parse-σ, between positional and general faith again derives the right ranking for this intermediate stage:

- 54) Max[Seg]-σ1 >> Parse-σ >> Max[Seg]

- 55) Parse-σ rules out unfooted syllables... except those protected by Max(Seg)-σ1

/yu_ru.ná.ca/	Max[Seg]-σ1	Parse-σ	Max[Seg]
yu_ru.(ná.ca)		**!	
yu.(ná.ca)		*	*
[(ná.ca)]	*!		**

2.3.4 Faithfulness to morphological roots

The example of a Specific Faith stage relativized to roots that I present here is somewhat different from the rest in that it comes from historical sound change. The

²⁵ Admittedly, since these authors do not provide percentages of outputs that conform to each particular pattern, we do not know how much of a stage this really was. However, it seems encouraging at least that the two children I've used here used this truncation pattern on different words, at the same age.

relevant data, raised by Albright (to appear), come from a change from Middle High German (MHG) to Modern Northeast Yiddish, in which the process of final obstruent devoicing disappeared.

In Middle High German, final devoicing held in both roots and affixes, so that forms were either voiceless across the board, or alternating:

- 56) *Source language ranking (MHG):*
 *FinalVcdObs >> Ident[vce]-Rt, Ident[vce]

In the case of roots, final consonants that had alternated in MHG between voiced and voiceless became uniformly *voiced* in Yiddish: compare the MHG nominative singulars and plurals in 57a) below with the corresponding Yiddish forms in 57b) (data from Albright, to appear examples 4 and 5) (data cited from Katz, 1987):

- 57) *MHG root-final voiced obstruents became voiced in Yiddish*

a) Middle High German forms ²⁶		b) Yiddish forms		glosses
nom. sing.	nom. plur.	sing.	plur.	
lop	lobe	loyb	loyben	'praise'
rat	reder	rød	reder	'wheel'
tak	tage	tøg	teg	'day'
hus	hiuzer	hoyz	hoyzer	'house'
brief	brieve	briv	briv	'letter'

At the same time, Albright points out that the two Middle High German affixes that similarly alternated became uniformly *voiceless* in Yiddish:

²⁶ This may not be the correct phonetic transcription of the vowels for MHG forms, but the final obstruents in the singular forms are definitely correct.

58) *MHG affix-final voiced obstruents became voiceless in Yiddish*

a) MHG affix	b) Yiddish affix	c) Yiddish examples	gloss
[-ik, -ige] (adjectival suffix)	[-ik, -ike]	lebedik, lebedike, lebedikən, lebediker	‘lively’
[ap, ab, abe] (preposition/ prefix)	[ɔp]	ɔpesn	‘eat up’

Thus it would appear that in the transition between Middle High German and Yiddish, speakers re-ranked from the fully unmarked grammar in 56) to the Specific-F ranking in

59):

59) *Post-sound change ranking (Yiddish) – c.f. 56):*
Ident[vce]-Rt >> *FinalVcdObs >> Ident[vce]

The explanation for why root-final consonants all became *voiced* rather than voiceless is a somewhat separate matter; see Albright (to appear) for the argument that it was paradigmatic leveling to the plural, and for a somewhat different approach to this data in general. But as for this re-ranking of *FinalVcdObs, Albright says:

60) “The older stage of the language provided no evidence for the relative ranking of Ident-IOLexCat(voi)²⁷ and Ident-IO(voi) [...] Therefore, we have no particular reason to expect that a demotion of the ban on voiced codas should have placed it below one constraint but not the other.” (Albright, to appear: 9)

My goal is to suggest a mechanism whereby it *is* predicted that learners who are abandoning the fully M >> F ranking in 56) first move onto an intermediate ranking like 59). However – as Albright points out in his footnote 8: “If some external force managed

²⁷ His version of Ident[vce]-Rt

to create voiced obstruents just at the end of stems, but not affixes, [a ranking bias of Spec-F >> Gen-Faith] would indeed learn exactly the right grammar (attested in Modern Yiddish.) However, [that bias] does not straightforwardly explain why voiced codas should be created in the first place.” It is certainly true that the extent to which my approach to gradual re-ranking can extend from synchronic developmental stages to diachronic sound change is not known, and will not be pursued further in this work.²⁸

2.4 Summary of the data

The data presentation of this section has pushed a particular view of children’s intermediate stages. Starting from the broad observation that children’s grammars increase in markedness as they develop, I have suggested that the grammars of many such stages can be characterized with rankings in which a specific version of a constraint is crucially ranked above its more general counterpart(s), either markedness or faithfulness:

- 61) *Intermediate stages*
a) The Specific-M stage: Specific-M >> Faith >> General-M
b) The Specific-F stage: Specific-F >> Markedness >> General-F

With this view of the data, I now turn to the question of how a BCD learner might be reliably coerced into passing through such stages.

²⁸ The crucial obstacle for applying this dissertation’s model in any discussion of sound change would be finding the *trigger* for re-ranking is, since this trigger cannot be the overt observational errors that child learners have at hand. Albright (to appear) contends that the relevant mechanism of sound change in this case was the learners’ decision to use the plural as the paradigm’s base; how this approach and my own might be integrated is a very interesting question for future research.

3. The theory of intermediate stages: Error-Selective BCD

As previewed in section 1, the heart of my proposal is that the learner should still use BCD as their re-ranking algorithm, in all its over-efficient splendour, but be conservative as to which errors it allows BCD to see when doing its re-ranking. Below I provide this mechanism, illustrate how it works using some of the examples we've already seen, and provide some discussion of its assumptions and workings.

A side note about an alternative theory before I begin: the OT phonological acquisition literature has seen much recent work using the Gradual Learning Algorithm (Boersma, 1997 *et seq.*), which is inherently designed to go through stages of acquisition. I alluded in chapter 1 to some problems that the GLA has in finding correct end-state grammars – all of these were to do with its choice not to retain errors for later reasoning, as the BCD does with its Support. Chapter 3 will turn to the GLA in earnest, and to the kinds of attested intermediate stages that it can (and cannot) derive.

3.1 The Error-Selective Learning proposal

Compared to straight BCD, my Error-Selective Learning approach (ESL) is different in two key respects: (a) what it does when it makes an error, and (b) what it does when it learns. With respect to the first, ESL retains the notion of the Support as the repository of errors that have been learned from, and which will be kept in mind each time re-ranking takes place. But ESL also uses another storage facility, called the Error Cache, which acts a holding pen for all the ERCs made on-line by the current grammar. Making an error does not trigger learning, but rather just an update of the Cache.

Periodically, the learner is triggered to stop merely accumulating errors and actually learn a new ranking. This triggering is done by a particular markedness constraint that has assigned Ls to ERC rows in the Cache – the details of how a constraint triggers learning are tied to input frequencies in a way discussed below. With respect to the second difference: learning proceeds in two steps. The first step is error selection: one particular error is chosen from the Cache to be added to the Support, and the second step is just re-ranking using the BCD algorithm already adopted.

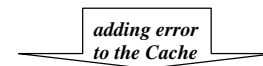
3.1.1 What happens when an error is made: a Specific-M example

Error-Selective learning starts, as with normal BCD, with the making of errors and the building of ERC rows. But unlike the straightforward BCD approach: errors don't immediately appear in the Learning Support once they're made; nor do they trigger re-ranking. Instead, when an error is made its resulting ERC row is added to a temporary storage area, which I will call the Error Cache:

62) *How the Error-Selective BCD learner responds to making an error*

a) At an Early Stage: this error is made...

	/tost/	NoCoda	*CompCoda	Max-
	tost	*!	*	
	tos	*!		*
	to			**



b) ... the learner adds it to the current Error Cache:

Input	Winner ~Loser	NoCoda	*CompCoda	Max-Seg
/tost/	tost ~ to	L	L	W
/piz/	piz ~ pi	L	e	W

c) ... but the Learning Support Table is NOT updated:

Input	Winner ~ Loser	NoCoda	*CompCoda	Max- Seg
... empty, waiting...				

And once the error has been added to the Cache: nothing else happens. While the learner has noted they've made an error, they do nothing immediate in response to that error, and so continue to use their current grammar, continue to make these (and other) errors and add them to the Cache, and the Cache keeps growing, until:

3.1.2 How learning is triggered

In Error-Selective learning, re-ranking is triggered when a constraint overcomes the *Violation Threshold* – that is, when some constraint has assigned Ls to more than x words in the Error Cache. I will refer to this particularly offending constraint as the *Trigger Constraint*, because it has triggered learning. If e.g. the violation threshold is 3, then as soon as some constraint assign an L to three different winner~loser pairs in the Error Cache, learning is triggered:

63) *a (sample of an) Error Cache that triggers learning:*

Input	Winner ~Loser	NoCoda	*CompCoda	Max	*CompOnset
i) /frend/	frend ~ fe	L	L	W	L
ii) /piz/	piz ~ pi	L	e	W	e
iii) /greᵖ/	greᵖ ~ ge	L	L	W	L
iv) /ti/	ti ~ si	e	e	e	e

This Error Cache already shows the benefit of defining *ComplexCoda as a less stringent version of NoCoda. An L assigned by *ComplexCoda is always accompanied by an L from NoCoda, as in the errors 63i) and 63iii) above; whereas NoCoda can assign

Ls when *ComplexCoda is indifferent, as in 63ii). As a result, a more stringent markedness constraint will always reach its Violation Threshold before a less stringent one (much more on this point in section 4.)

3.1.3 Step 1: Choosing an error to learn from

Once learning has been triggered, the Error-Selective learner must choose one error from the cache – the Best Error – to learn from. As a first pass, we can choose errors via the two criteria I give below:

64) *The Error Selection Algorithm (ESA) (first pass)*

Choose as the best error that row in the Cache which:

- a) has an L assigned by the Trigger Constraint, and of those, the one that
- b) has the fewest Ls assigned by *other* Markedness constraints,

I will return to the general consideration of why the ESA looks the way it does in section 3.2 below: for now, let us just apply these criteria to the Cache in 63) above. Criterion (a) eliminates 'tea', because it does not have a NoCoda violation. Criterion (b) eliminates 'friend' and 'grape', because they also have Ls assigned by *ComplexCoda and *CompOnset. Thus, the Best Error of these four is 63ii):

65) *piz ~ pi is the chosen Best Error*

Input	Winner ~Loser	NoCoda	*CompCoda	Max	*CompOnset
i) /frend/	frend ~ fe	L	L	W	L
ii) /piz/	piz ~ pi	L	e	W	e
iii) /greᵖ/	greᵖ ~ ge	L	L	W	L
iv) /ti/	ti ~ si	e	e	W	e

After choosing the best error, the learner adds that error to the Support and empties the Cache. To illustrate how far the learner has gotten now:

66) How the Error-Selective BCD Learner gets from errors to a new Support

a) At an Early Stage: this error is made...

Input	Winner ~Loser	No Coda	*Comp Coda	Max
/frend/	frend ~ fe	L	L	W
fend	*f	*	*	*
fēn	*f	*	*	*
fe				**

error added
to the Cache

b) NoCoda becomes the trigger constraint

Input	Winner ~Loser	No Coda	*Comp Coda	Max
/frend/	frend ~ fe	L	L	W
/piz/	piz ~ pi	L	e	W
/greʔp/	greʔp ~ ge	L	L	W
/ti/	ti ~ si	e	e	e

step 1:
ESA

d) Error Cache cleared...

Input	Winner ~Loser	No Coda	*Comp Coda	Max
... empty, waiting ...				

c) the pre-existing Learning Support Table

Input	Winner ~Loser	No Coda	*Comp Coda	Max
... empty, waiting...				

e) ... and Support updated

Input	Winner ~Loser	No Coda	*Comp Coda	Max
/piz/	piz ~ pi	L	e	W

And now the learner moves on to step two of learning, which is simply:

3.1.4 Step 2: Applying BCD

In the case above, this means learning from the new Support piece of data in 66e) above. As we have seen a few times already, this ERC demonstrates the need to demote NoCoda below Max, but the e assigned by *Complex Coda means that the BCD learner can install this last constraint at the top of the hierarchy. Thus, ranking from this Support will get us to the ranking *ComplexCoda >> Max >> NoCoda – and this is the Specific-

M ranking we saw in section 1.2. This grammar protects the markedness of singleton codas, because the error ‘peas’ in the Support demonstrates that the target language tolerates them, but it still reduces complex codas to singletons.

3.1.5 A second example: a Specific-F stage

To see the role of Faithfulness in this error-selective decision-making process, I return to the French example of complex onsets in stressed vs. unstressed syllables. Recall that at the first stage of data, repeated below using Clara’s data, all onset clusters are reduced:

67) Clara’s stage 1 – all complex onsets reduced to singletons (1;0.28-1;09.01)

stressed syllables			unstressed syllables		
Target	Child	Gloss	Target	Child	Gloss
/kʁa.'kʁa/	[ka.'kæ]	‘Cracra’ (name)	/bʁi'ze/	[bœ'çi:]	‘broken’
/plœs/	[pœ:]	‘(s/he) cries’	/apʁi'ko/	[pupæ'ko]	‘apricot’
/flœs/	[βœ:]	‘flower’			

The state of Clara’s learning during stage one – just before stage two – is reflected in the Error Cache below:

68)

Input	Winner ~Loser	*Complex Onset	Max(Seg)	Max(Seg)-σ'
i) /plœs/	plœs ~ pœ:	L	W	W
ii) /flœs/	flœs ~ βœ:	L	W	W
iii) /kʁa.'kʁa/	kʁa.'kʁa ~ ka.'kæ	L	W	W
iv) /bʁi'ze/	bʁi'ze ~ bœ'çi:	L	W	e
v) /apʁi'ko/	apʁi'ko ~ pupæ'ko	L	W	e

This Error Cache has a number of errors with violations of *Complex – so, let us imagine we are at the stage where *Complex has triggered learning. Assuming that we have narrowed the candidates to these five using criterion (b) of the ESA, we now need a way of choosing among the errors in 68), and for that we need a third criterion:

69) *The Error Selection Algorithm (ESA) (full version)*

Choose as the best error that row in the Cache which:

- a) has an L assigned by the Trigger Constraint, and of those, the one that
- b) has the fewest Ls assigned by *other* Markedness constraints, and of those the one that
- c) has the most Ws assigned by other *Faithfulness* constraints

Criteria (c) tells us that of the errors in 68), we want an error that has the *most* Ws among the faithfulness constraints. Given this Cache, this will mean one that has Ws in both Max-Seg and Max-σ' columns, e.g. one of the first two:

70) *The best error(s) chosen*

Input	Winner ~Loser	*Complex Onset	Max(Seg)	Max(Seg)-σ'
i) /p œɛ/	'plœɛ ~ pœ:	L	W	W
ii) /ʃ œɛ/	'ʃlœɛ ~ βœ:	L	W	W
iii) /kχa.'kχa/	kχa.'kχa ~ ka.'kæ	L	W	W
iv) /bɛi'ze/	bɛi'ze ~ bæ'çi:	L	W	e
v) /abɛi'ko/	abɛi'ko ~ pupæ'ko	L	W	e

And adding one of these two ERCs to the Support will allow the specific >> general IO-faith bias to install just Max(Seg)-σ' above *ComplexOnset, giving us our intermediate stage:

71) *The Support*

Input	Winner ~ Loser	*Complex Onset	Max-Seg	Max-Seg (σ')
i) /p œɛ/	'plœɛ ~ pœ:	L	W	W

72) *Resulting BCD ranking*

Max[Seg]-σ' >> *ComplexOnset >> Max[Seg]

3.2 Discussion of the ESA, and Error-Selective Learning more generally

ESL is a way to learn from errors that will change the grammar as minimally as possible – while still being able to use the restrictive power of BCD. Since errors in the cache were all created by a current ranking, the ESA can find errors that only require small revisions to the present grammar (i.e. the demotion of a small number of L-preferring Markedness constraints), given the current lexicon.²⁹

3.2.1 Analyzing the three ESA criteria for choosing errors

The Error-Selection Algorithm is the mechanism that decides which stages the learner goes through, because it chooses the errors that BCD builds its grammars from. The ESA's three criteria are thus built to ensure that the learner will choose errors that derive the kinds of intermediate stages discussed above. They are based on the logic of ERC rows – what their Ws and Ls that they contain tell us – and they are arranged in order of decreasing importance.

²⁹ As Elan Dresher points out (p.c.) the Error-Selective learner's reliance on the lexicon to provide minimal changes to the grammar is reminiscent of the Triggered Learning Algorithm (Gibson and Wexler, 1994). In this way, ESL is also similar, perhaps more so, to the learning model of Fodor (1998a,b) and subsequent – thanks to Lyn Frazier for discussion.

To understand the first two criteria: recall that the Markedness constraints that assign Ls to winner-loser pairs are those which are currently ranked too high in the current grammar; they are in some way responsible for the particular error that has been made.

The first criterion of the ESA is that chosen errors must have an L assigned by the Violation Threshold. This means that the learner will attend to the most frequent marked structure in the target that its current grammar does not allow – the constraint that triggered learning in the first place. The second criterion is that the chosen error has as few *other* Ls assigned by Markedness constraints. This means that the learner will choose an error that makes BCD demote as few other Markedness constraints as possible – in other words, that it will teach it as few new things as possible.

Together, these two criteria derive Specific-M stages, like the singleton vs. complex codas in §3.1.1-4 above. Once NoCoda has overcome the VT, the learner will add an error to the Support that shows that NoCoda must be demoted, but which says nothing about the ranking of as many other Markedness constraints as possible (like *ComplexCoda.) Given the BCD bias for high-ranking Markedness constraints, choosing an error that does not prove the need for demoting *ComplexCoda will allow it to stay at the top of the hierarchy.

To understand the third criterion: recall that among the faithfulness constraints that assign Ws in an ERC row, at least *one* of them must be ranked higher in the target grammar than in the current one. It is also important to realize the W-assigning properties of more vs. less stringent faithfulness constraints. A marked structure that is assigned an L by a less stringent faith constraint will also get one from the more stringent, general

faith constraint – e.g., if an onset is voiced in the target winner but devoiced in the loser, it will receive Ws from both Ident[vce]-Ons as well as general Ident[vce]. This in turn means that errors in which a marked structure appears in a privileged position will be assigned *more* Ws than if the same marked structure appears in a less privileged position – i.e., that compared to the voiced obstruent in onset that got two Ws assigned by Id[voice] constraints above, an obstruent in coda position that is devoiced in the loser will only garner a W from the general Id[voice] constraint.

The third criterion of the ESA is that chosen errors should have as *many* Ws assigned by faithfulness constraints as possible. As we've just seen, this criterion will choose errors that have marked structures in privileged contexts.

In conjunction with a ranking bias for specific >> general IO faith, this third criterion derives Specific-F stages; this was illustrated in the complex onset example of §3.1.5. Recall that BCD installs as few IO-faith constraints as it can and still resolve its errors, and that its biases ensure that it tries installing specific faithfulness constraints before general ones. So by choosing errors with Ws assigned by very specific faithfulness constraints, the learner ensures that BCD builds a grammar that allows marked structures only in those very specific contexts, and gradually learns their true scope as more errors are added to the Support.

3.2.2. Terminating ESL and converging on the end stage grammar³⁰

How can we be sure that Error-Selective Learning will terminate? To do so, we must make sure that the final time an error is added to the Support, it is the *ONLY* remaining error – that is, that nothing unexplained remains in the Cache. With the system

³⁰ Thanks to John McCarthy for alerting me to this very necessary aspect of the proposal.

as it stands, it is unfortunately the case that some errors will *never* get added to the Support. If after some late stage of learning, the number of lexical items in the language on which the learner is still making errors *is less than the Violation Threshold*, then learning will never be triggered again but yet the learner will not have reached the end state. This is not a welcome result.

To make sure the Error-Selective learner terminates, we must make sure that eventually even a single error in the Cache can be added to the Support. Thus, I propose that the Violation Threshold is not just a fixed number, but rather a value that changes over time. The VT will begin fixed at its highest point, and decrease over time until it is eventually at one.³¹

With this caveat, the Error-Selective Learner will eventually end up with the same grammar as straight BCD would. Every time a constraint exceeds the Violation Threshold, some new error is chosen to add to the Support, and once BCD learns from an error, it is never made again. Because the Error Cache is emptied every time a new error is added to the Support – this will prevent the learner from being trapped in being triggered from very frequently violated constraints over and over again.³² Eventually all the errors necessary to finding the correct grammar will be added to the Support, at which point no more errors are made and the learner will have reached the final state.

³¹ Two remarks. First, it might be fruitful in discussing language evolution to consider the effects of not letting the VT get as small as one, as a way to quantify how generalizations for which there is infrequent evidence (from perhaps only a few lexical items) are lost over generations of language acquisition. Second, note that the idea of a decreasing VT is related, but not directly analogous, to the decreasing re-ranking plasticities of the GLA – see chapter 3.

³² Although depending on the repairs that learner choose, Trigger Constraints may continue to be violated in later Error Caches. See section 3.3 below.

3.2.3 Irrelevant markedness violations

In choosing an error to move from Cache to Support as defined in the ESA above, the learner is taking into consideration *all* Markedness violations, and this may have somewhat unanticipated consequences. I illustrate this point in 73) below, with a slightly more articulated Error Cache involving NoCoda and *ComplexCoda. From this Cache, the learner will fail to go through the intermediate stage of singleton codas – just because the error with the singleton coda has a somewhat marked coda consonant, while another error has a complex coda with less marked segments:

73) *an Error Cache that triggers learning – but of a complex coda:*

<i>Input</i>	<i>Winner ~Loser</i>	<i>NoCoda</i>	<i>*Comp Coda</i>	<i>Max</i>	<i>*Comp Onset</i>	<i>*Fricative</i>	<i>*VcdObs</i>
i) /frend/	frend ~ fe	L	L	W	L	e	e
ii) /piz/	piz ~ pi	L	e	W	e	L	L
iii) /gre'p/	gre'p ~ ge	L	L	W	L	e	e
iv) /ti/	ti ~ si	e	e	e	e	e	e
v) /pant/	pant ~ pa	L	L	W	e	e	e

By counting the Ws assigned by markedness constraints other than NoCoda, we can see that criterion (b) of the Error-Selection Algorithm will choose the ERC row in 73v) with its one other L-prefering markedness constraint over any of the others, including 73ii) with its marked coda [z].

The upshot is that whether or not an ESL learner goes through the singleton coda stage depends on how marked the singleton vs. complex codas are in the particular errors in the Cache. So, while the learner will always be triggered to learn by NoCoda before *ComplexCoda, they are *not* guaranteed to choose a best error that will push them

through this stage. Overall, this seems like the right prediction: e.g. the onset sonority effects that Trevor and Julia display are not true of all learners. With respect to stages of specific faithfulness, recall the conflicting results of Rose (2000) and Kehoe and Debove-Hilaire (2003) as to which French onset clusters give rise to the intermediate stage of stressed-syllable faith only in different children. One could complicate the learner further to ensure the stringency result – e.g. by adding more analysis of the other M violations – but it is not clear that the current, simpler method is necessarily undesirable.³³

3.2.4 Choosing among positional faithfulness contexts

One aspect of criterion (c) – which favours errors with as many faithfulness Ws as possible – comes from French complex onsets example. With the right set of faithfulness constraints, triggering the ESA with the markedness constraint *ComplexOnset should lead the learner to pick an error that has a complex onset in a monosyllabic word. In a one syllable word, the syllable with the Markedness violation will be in both the stressed and initial syllables, and this will result in the most faithfulness Ws:

74) *A French learner's Error Cache, repeated*

Input	Winner ~Loser	*Complex Onset	Max-Seg	Max-Seg (Stressed σ)	Max-Seg (σ 1)
i) /plœs/	'plœs ~ pœ:	L	W	W	W
ii) /flœs/	'flœs ~ βœ:	L	W	W	W
iii) /kʁa.'kʁa/	kʁa.'kʁa ~ ka.'kœ	L	W	W	e
iv) /bʁi.'ze/	bʁi.'ze ~ bœ.'çi:	L	W	e	W
v) /abʁi.'ko/	abʁi.'ko ~ pupœ.'ko	L	W	e	e

³³ John McCarthy (p.c.) points out that Error Caches like 73) will not arise only if two conditions are met: (i) the Violation Threshold is (initially) set sufficiently high to give the learner a representative sample of errors, and (ii) languages are assumed to be harmonically complete (on this notion, see esp. Smolensky and Legendre (2006) chapter 14.)

It will then be up to BCD to decide which faithfulness constraint to install above *ComplexOnset; let us assume that this French learner has already established via an illustrative Context Table from chapter 1 that French's initial syllables and stressed syllables are in no subset relationship. In the absence of any other relevant data, the learner may well choose to install Max(Seg)- σ_1 over *ComplexOnset. While this has no deleterious consequences for the end-state grammar, it does predict that French-learning children could go through a stage where onsets are retained only in initial syllables, and nowhere else.

3.2.5 The Violation Threshold and extra-grammatical factors

In the Error-Selective model, Violation Thresholds provide the interface between the language-specific module, with its knowledge of phonological constraints and abstract representations, and all more general cognitive factors in language development.

As already mentioned in the previous section, learners' VTs must decrease over time to ensure that they can eventually get all necessary errors in the Support to finish learning. The *rate* at which the VT decreases will determine the speed with which intermediate stages are overcome and new grammars are learned – and the correct initial values and rate of decrease are empirical questions (which this dissertation has far too little data to answer.) But it seems plausible that both of these parameters would vary from child to child as a function of individual cognitive abilities, and from context to context as a function of all other current cognitive demands on a child.

Allowing for different Violation Threshold values at different moments in learning also opens the ESL to one possible treatment of variability between rankings

over the course of development -- and even perhaps regressions, where learners temporarily return to an earlier stage after having mastered a later one. These possibilities will be the focus of section 5.

3.3 Illustrating ESL: a case study of Trevor and Julia's onset clusters

This section dissects Trevor and Julia's stop-initial onset cluster acquisition from §2.2.2 in more detail, with the ESL proposal in mind. I present a few stages along the way to complete mastery of onset clusters, and demonstrate how the Error-Selective learner can get from each stage to the next. Note that I follow each child up until the point where their data becomes insufficiently transcribed, meaning that that they are both not fully finished cluster acquisition at their final stages here.

I focus here just on stop-initial and s-initial clusters, because they are sufficiently attested in the data to make what I think are confident generalizations. While the sonority constraints that I adopt make predictions about the concurrent acquisition of other fricative-initial onset clusters, I leave them out of the current analyses.³⁴

3.3.1 Trevor

This section will focus on Trevor's first three stages of onset cluster acquisition. At his first stage, which lasts until approximately 2;2, all clusters are reduced to singleton onsets only. At his second stage, which lasts about two months from 2;3-2;4, he permits

³⁴ It should perhaps also be noted, however, that the sonority difference between fricatives and stops may not in fact be relevant to the typology of permissible onset clusters: see Morelli (1999). If this were true, then the sonority hierarchy relevant to building the set of Onset Sonority Constraints in 27) would be collapsed to contain as its less sonorous element 'obstruents' – and this would make different predictions about the stages discussed here.

stop-r and stop-w clusters³⁵ but continues to reduce stop-liquid and all s-initial clusters to singletons. His third stage begins around 2;5, when he adds stop-liquid clusters to his inventory, but still reduces s-initial ones. To summarize, then (using capital S for stops):

- 75) Trevor's onsets at the first three stages
- | | | |
|----------|---------------------------------------|---------------------|
| Stage 1: | [CV...], | *[CCV...] |
| Stage 2: | [SrV...], [SwV...], [SV...] | *[SIV...], [sCV...] |
| Stage 3: | [SIV...], [SrV...], [SwV...], [CV...] | *[sCV...] |

An important caveat: these stages are in fact abstractions from the quantitative patterns of cluster preservation and reduction that Trevor passes through. The numbers I provide below will show that these are the *prevalent* productions at each stage, but clusters of each type are, in fact, reduced and retained to varying degrees throughout this period. For now I put aside the treatment of this variation, and discuss the ESL route through Trevor's stages as though they were all categorical, with the promissory note that section 5 will deal with some of these variation issues.

Getting from stage 1 to stage 2:

We begin at stage one:

³⁵ Trevor has only one stop-j cluster – 'piano' – which is consistently reduced. Perhaps Trevor is not perceiving this glide, or perhaps independent markedness constraints are ruling out [pj].

76) The end of Trevor's stage 1: all clusters reduced (up to 2;2)

age	output	raw #				percentages			
		stop-l	stop-r	tr	sC	stop-l	stop-r	tr	sC
2;0	C...	31	31	32	15	100.0	86.1	82.1	100.0
	CC...		5	7		0.0	14.7	23.3	0.0
2;1	C...	65	29	23		94.2	69.0	76.7	
	CC...	4	13	7	3 ³⁶	5.8	31.0	23.3	
2;2	C...	30	34	19	32	100.0	82.9	50.0	100.0
	CC...		7	19		0.0	17.1	50.0 ³⁷	0.0

The relevant fragment of Trevor's grammar at this stage is fully M >> F: all of the Onset

Sonority constraints (*TW, and everything else) rank above faithfulness, e.g.:

77) *TW, *TR, *TL >> Max

Based on the words in 69) above, Trevor's Error Cache at 2;2 includes errors as in 78)

below:

78) *A fragment of Trevor's Error Cache at the end of stage 1*

Word	Winner-Loser	*TW	*TR	*TL	Other Mkdness	Max
'blocks'	blaks ~ gak	L	L	L	?	W
'glasses'	glæstɪz ~ gæfɪʃ	L	L	L	?	W
'clock'	klak ~ kak	L	L	L	?	W
'cracker'	kræki ~ kaka	L	L	e	?	W
'train'	tre'n ~ te'n	L	L	e	?	W
'between' ³⁸	bə'twɪn ~ ti:	L	e	e	?	W

³⁶ I consider these three tokens of /sC/ productions to be an aberration.

³⁷ As mentioned in section 2, Trevor's acquisition of [tr] is different from the rest of his obstruent-r clusters.

³⁸ This error is not from the corpus, but is rather inferred – as the footnote below points out, Trevor had so few stop-w inputs that we can't be sure when he started allowing them. Whether it was at 2;2 or earlier, he would have initially not tolerated them and so made errors such as this one.

To trigger learning, one of these markedness constraints must overcome the Violation Threshold. Since these markedness constraints are in a stringency relation, the first constraint to do so will be the most stringent, i.e. *TW (already shaded in the Cache above – see below on why I've shaded *TR as well))

Once *TW has triggered learning, Trevor now must search the Cache to find the best error, which must at least (a) violate *TW and (b) violate as few other M constraints as possible. Since Trevor's second stage permits both stop-w and stop-r clusters, one possibility is that among the errors violating *TW, the best one also violated *TR – that is, that those few errors with stop-w clusters happened to have more *other* Markedness Ls than the Cr ones. A second possibility is that between what I have called stages 1 and 2, Trevor learns twice in quick succession: triggered first by *TW, and then triggered again by *TR almost immediately afterwards. This is not so implausible given that Trevor has very few words with stop-w clusters compared to stop-r words – in fact, I have not given numerical data from Trevor's stop-w clusters above precisely because they are so infrequent.³⁹ This suggests that *TR would have reached its threshold soon after *TW. (Drawing the line between the activity of these two constraints is in any case difficult given that at this stage /r/ is frequently mapped to [w]).

In either case: once the learning dust settles, Trevor's Error Cache has been cleared, and his Support has been updated with an error containing a Cr cluster. From this Cache, I have chosen 'train':

³⁹ In the entire corpus up until 2;2, his only two attempted stop-w clusters are [twi:] 'between' (2;1.14), and [skɪz] 'squeeze' (2;1.14).

79) *Trevor's Support after step 1 of learning:*

Input	Winner ~Loser	*TW	*TR	*TL	Max
'train'	tre'n ~ te'n	L	L	e	W

And after step 2 – applying BCD to this Support – Trevor gets the ranking in 78) below that protects Cr and Cw clusters:

80) *Stage 2 ranking, from the Support in 79)*
 *TL, *... >> Max[Seg] >> *TW *TR
newly demoted

Getting from stage 2 to stage 3:

At stage 2, Trevor's ranking allows stop-r onsets, but still reduces most other onset clusters:

81) *Trevor's stage 2 (numbers by output):*

age	output	raw #				percentages			
		stop-l	stop-r	tr	sC	stop-l	stop-r	tr	sC
2;3	C...	23	12	35	17	69.7	31.6	89.7	77.3
	CC...	10	26	4	6	30.3	68.4	10.3	22.7
2;4	C...	28	18	25	16	62.2	41.9	53.2	84.2
	CC...	17	25	22	3	37.8	58.1	46.8	15.8
2;5	C...	14	4	5	21	35.9	20.0	22.7	72.4
	CC...	25	16	17	8	64.1	80.0	77.3	27.6

After a couple months of this, Trevor is again triggered to learn by an error in his Cache. Since his Cache was cleared when he added an error to the Support to get to stage 2, none of his errors on stop-w and stop-r clusters remain there. In the present errors (see the Cache in 82 below), *TL prefers the losers in his onset cluster errors – but so do *TW

and *TR. Thus, all *three* of these constraints will overcome the VT and trigger learning at the same time – I have only chosen *TL to shade as the Trigger Constraint below because it is this constraint whose position will be changed in the eventual re-ranking:

82) *A fragment of Trevor's Error Cache at 2;5*⁴⁰

Word	Winner ~Loser	*TW	*TR	*TL	*SN	*ST	Other Mkdness	Max
'glass'	glæs ~ gæs	L	L	L	e	e	?	W
'play'	ple' ~ pe'	L	L	L	e	e	?	W
'cleaner'	klɪnɪ ~ kɪ:nə	L	L	L	e	e	?	W
'sneakers'	snɪkɪz ~ ɔnikəθ	L	L	L	L	e	?	W
'stick'	stɪk ~ dɪk	L	L	L	L	L	?	W

An important question about this stage is why, if *TW and *TR have been demoted, clusters like TL are still being deleted rather than mapped to better sonority clusters: in other words, why does 'glass' come out as 'gas' and not 'gwas' or 'gras'? The explanation will have to come from the ranking of other constraints. As we've seen, deletion (e.g. /gl/ → [g]) violates Max; a featural mutation like (/gl/ → [gw]) violates Ident constraints, as well (possibly) as other markedness constraints; some of these latter constraints must be currently ranked above Max in Trevor's grammar.⁴¹

Getting back to the Cache in 82) above): Trevor will now choose a Best Error that violates *TL but no *more* onset cluster constraints. Supposing that he chooses his error on 'glass', he adds it to the Support, and applies BCD to find the new ranking in 84):

⁴⁰ Beyond the data cited earlier – the new errors, 'sneakers' and 'stick' are from 2;4.3.

⁴¹ Beyond these facts about Trevor, it is more generally the case that children's grammars often prefer deletion over other repairs. I cannot fully treat this issue here, but it is an interesting and outstanding question to what extent the deletion preference can be reduced to the activity of other constraints or requires e.g. some initial rankings among faithfulness constraints.

83) *Trevor's new Support:*

Input	Winner ~Loser	*TW	*TR	*TL	*SN	*ST	Max
'train'	tre'n ~ te:n	L	L	e	e	e	W
'glass'	glæs ~ gæs	L	L	L	e	e	W

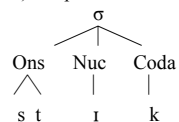
84) *Stage 3 ranking, from the support in 81):*

... *SN, *ST ... >> Max >> *TW, *TR, *TL
newly demoted

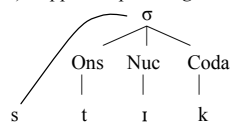
At this point, the major class of onset clusters that Trevor has not learned is the s-initial cluster, and the best treatment of this lag depends on assumptions about Trevor's syllabification of those inputs with sC sequences. The issue is whether the child is treating them as true sC onset clusters, or as singleton C onsets with the s in an adjunct position outside the syllable. (In the developing English context, see e.g. Barlow 2001, Goad and Rose, 2004; Chambless, 2006):

85) *Two possible syllabifications for 'stick'*

a) *complex onset*



b) *s-appendix plus singleton onset*



Trevor's consistent pattern of /sC/ → [C] reduction, regardless of relative sonority, allows two possible interpretations. The assumption I've been making above so far is that he is treating them as clusters, and so they violate so many sonority sequencing constraints that the ESA will not yet have chosen to add them to the Support (recall that the ESA's criterion (b) wants errors that have as few L-preferring Markedness constraints

as possible, ignoring the Trigger Constraint.) The alternative hypothesis is that they are adjuncts and the *Appendix constraint has yet to be demoted, as below.

86) *An alternative version of the Error Cache at 2;5*

Word	Winner ~Loser	*TW	*TR	*TL	*Appendix	Other Mkdness	Max
'glass'	glæs ~ gæs	L	L	L	e	?	W
'play'	plɛ' ~ pɛ'	L	L	L	e	?	W
'cleaner'	klɪnɪ ~ klɪnə	L	L	L	e	?	W
'sneakers'	sni:kɪz ~ əni:kəθ	e	e	e	L	?	W
'stick'	stɪk ~ dɪk	e	e	e	L	?	W

3.3.2 **Julia**

Julia's acquisition of onset clusters is a little more complicated than Trevor's, because she treats s-initial clusters differently depending on their sonority profile. But her first two stages are just like Trevor's (see below), so I pick up her ESL story at stage 2:

87) *Stage 1: singleton onsets only (up to 1;9)*

Stage 2: stop-r and stop-w onsets only (during 1;10)

Starting with stage 2:

As with Trevor, Julia's stage 2 reduces permits few s-initial onset clusters of any sort:

88) Julia's stage 2 (numbers by output):

age	output	raw #s			percentages		
		stop-r/w	stop-l	sC	stop-r/w	stop-r	sC
1;10	C...	9	21	10	26.5	95.5	80
	CC...	25	1	2	73.5	4.5	20

89) Representative data from Stage 2

clusters retained: stop-r, stop-glide			clusters reduced: stop-l, s-initial		
Target	Child	Age	Target	Child	Age
'drink'	[gwiŋk]	1;10.5	'spoon'	[pun]	1;10.8
'Grundy'	[gwani]	1;10.5	'sleep'	[sip]	1;10.7
'crackers'	[kwækəs]	1;10.14	'glasses'	[qaθəs]	1;10.10
			'please'	[pis]	1;10.10

Thus Julia's Support at stage 2 is equivalent to Trevor's – that is, it contains errors that demonstrate the need to demote *TW and *TR.

90) Julia's Support at stage 2:

Input	Winner ~Loser	*TW	*TR	*TL	*SN	*ST	Max
'cry'	krai~ kai	L	L	e	e	e	W

Getting from stage 2 to 3

Julia's stage 3 of learning, at around 1;11-2;0, adds s-stop and s-nasal clusters to her onset inventory – while stop-liquid, s-liquid and s-glide clusters continue to be reduced.

The table and data below illustrate this:

91) Stage 3: s+[-cont] (s-stop and s-nasal) onsets also appear at 1;11-2;0

age	output	raw #s				percentages			
		stop-r/w	stop-l	s-stop, s-nasal	sl, sw	stop-r/w	stop-l	s-stop, s-nasal	sl, sw
1;11	C...	22	1	4	5	100.0	97.6	14.3	80
	CC...	0	40	24	1	0.0	2.4	85.7	20
2;0	C...	23	3	1	7	100.0	92.1	3	87.5
	CC...	0	35	32	1	0.0	7.9	97	12.5

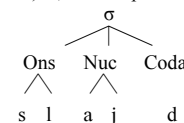
92) Representative data from Julia's stage 3:

clusters retained: stop-w, stop-r, s-stop, s-nasal			clusters reduced: stop-l, sl, sw		
Target	Child	Age	Target	Child	Age
'queen'	[gwin]	2;0.2	'clap'	[kæp]	1;11.15
'crown'	[kwaun]	2;0.2	'slipper'	[sipə]	2;0.18
'spilled'	[sprod]	1;11.22	'slide'	[sai:t]	1;11.16
'sneeze'	[snis]	1;11.26	'swim'	[fim]	1;11.15

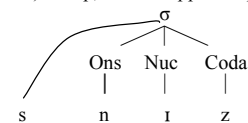
To get to this stage 3, Julia must choose a best error that will let her learn s[-cont] but not s[+cont] onsets. But what is the triggering constraint? It clearly cannot be onset sonority that prefers ST and SN clusters (look at the violations in the previous Error Cache.) Instead, I suggest that the right triggering constraint is *Appendix, and that the s-initial clusters that Julia learns at stage 3 are those that her grammar has parsed as containing not a complex onset but as a singleton onset and an s-appendix (see ranking below):

93) Julia's sC syllabifications

a) sl, sw: complex onset



b) s-stop, s-nasal: appendix plus singleton



Therefore, I adopt the two syllabification patterns in 93), to illustrate Julia's Error Cache below.⁴² With this assumption, we can see that if *Appendix triggers learning Julia will choose a best error with an s-[-cont] onset cluster:

94) *Julia's Error Cache at the end of 1;10*

Word	Winner ~Loser	*TW, *TR	*TL	*SL	*Appendix	Other Mkdness	Max
'please'	pliz ~ piz	L	L	e	e	?	W
'sleep'	slip ~ sip	L	L	L	e	?	W
'spoon'	sp ^h un ~ pun	e	e	e	L	?	W
'stairs'	sterz ~ deəz	e	e	e	L	?	W
'snake'	snek ~ nek	e	e	e	L	?	W

If for example Julia chooses 'spoon', her Support is duly updated as in 95), she then applies BCD, and she thus gets the stage 3 ranking:

95) *Julia's new Support*

Input	Winner ~Loser	*TW, *TR	*TL	*SL	*SN	*Appendix	Max
'cry'	kra ^l ~ kai	L	e	e	e	e	W
'spoon'	s.p ^h un ~ pun	e	e	e	e	L	W

96) *The new ranking at stage 3, from the Support in 93)*
 *TL... *SN, *ST... >> Max >> *TW, *TR, *Appendix
newly demoted

⁴² This split in her syllabification might be linked to her choice of segments when these clusters are reduced. As we have seen here, Julia reduces sl and sw clusters to [s], but s-stop and s-nasal clusters to the stop or nasal. If we assume that onset selection is driven by a preference for the least sonorous onset segment (with respect to children, see e.g. Gnanadesikan, 1995/2004; Pater and Barlow, 2003; Goad and Rose, 2004; cf. van der Pas, 2004), then it would be surprising that she chose to reduce complex onset /sn/ clusters to just the more sonorous nasal. If, however, that cluster is syllabified with only the nasal in onset, the pattern is explained.

An important point here is why SW and SL clusters don't surface faithfully at stage 3 by being parsed in the output as appendices. The answer must lie in the ranking of other markedness constraints: the likely candidates are syllable contact constraints. On the assumption that the appendix-onset boundary is assessed by such constraints, high-ranking constraints that prohibit too sharp a rise in sonority across the syllable boundary will rule out the appendix parse for SW and SL clusters (on syllable contact, see Murray and Venneman, 1983; Clements, 1990; on an OT analysis in the current system's spirit, see Gouskova, 2004):

97) *Julia's grammar chooses reduction of SL and SW*

'sleep' /slip/	Syllable Contact	*SL	Max-[Seg]	*Appendix
s.lip	*!			*
slip		*!		
^ɸ sip			*	*

3.3.3 Summary

This section has considered several stages in the acquisition of onset clusters by two different children. I have shown in ESL how the stringency of onset sonority constraints can predict well-attested stages, whereby better onset clusters are acquired before less good ones. I have also relied on s-initial clusters' variable syllabification to explain the differences between Trevor and Julia's development.

One interesting point of comparison between these two children is that Julia and Trevor go through the same kinds of onset cluster stages (modulo the sC differences), but Julia's are much *earlier* than Trevor's. Why should this be? In ESL there are two options: either she has more errors earlier, or she has lower Violation Thresholds. Both of these

options seem plausible, and while in the present system this is a mechanical, rather than an empirical issue, some important questions arise from the consideration of these mechanics.

The second option makes sense in the terms of §3.2.4 above, with which we can consider VTs as flexible thresholds of a rather psychological nature: affected by other cognitive demands and individual abilities, decreasing over time as language processing gets easier for the learner⁴³, and the like.

The first option – that Julia’s Error Caches grow faster than Trevor’s – raises the somewhat unanswered question of how precisely errors get made and into the Cache. The next section deals in part with this question, because it considers how frequency affects Error-Selective Learning. As we will see, what the empirical predictions connect are ambient lexical frequencies and constraint stringency on the one hand and *order* of acquisition on the other. However, they have nothing central to say about *rates* of acquisition. In any event, the difference between Julia and Trevor’s rate of onset cluster learning will remain outside the predictive domain of this fundamentally grammatical, not psychological, approach to acquisition.

4 The roles of frequency

4.1 The connection between frequency and Error-Selective Learning

There has been extensive discussion in the literature of how lexical frequency influences acquisition order. Drawing the right formal connections between frequency and grammar is clearly a long-standing point of controversy, no less tricky in the domain of acquisition than in phonological theory as a whole. However, an important fact that the

⁴³ Similar to the decreasing plasticity of the GLA – see chapter 3 section 1.

Error-Selective approach exploits is that incorporating frequencies or statistics across the lexicon into *learning* is logically independent from using those frequencies or statistics directly in the *grammar* – either in the definition of constraints or in the workings of EVAL.⁴⁴

In Error-Selective Learning, the ideas of Violation Thresholds and the ESA criterion (b) that favours errors with the fewest *other* Ls, together conspire to predict that order of acquisition should mirror markedness violation frequency. The more errors that a constraint assigns Ls to, the earlier one of those L-assigned errors will get into the Cache, and so the earlier it will be demoted.⁴⁵

In these two ways, the ESA only makes universal predictions about order of acquisition among more vs. less stringent M and F constraints (and then not even a completely deterministic prediction, as pointed out in §3.2.2). The relative re-ranking of constraints *not* in stringency relations, on the other hand, is in no way fixed beforehand. Instead, these ordering decisions will be specific to the target language, and the particular errors the learner has added to their Cache.

4.2 The connection between frequency and order of acquisition

The arguments in the literature connecting frequency and stages involve cross-linguistic comparisons, in both absolute and relative terms (for a recent brief review, see Beckman and Edwards, 2000). Here I discuss one robust example, from a series of studies that together demonstrate how differences in order of acquisition between

⁴⁴ Thanks to Sharon Goldwater for discussion of this point.

⁴⁵ On the Faithfulness side, the ESA criterion (c) that favours errors with the most Ws predicts something comparable but not hinged on frequency – that the more privileged positions a marked structure appears in, the earlier an error that forces its acquisition will be added to the Cache.

Germanic (German, English, Dutch) and Romance (Spanish, at least) are the result of language-specific input frequencies.

4.2.1 Data from cross-linguistic frequency: initial weak syllables vs. codas

The distillation of this cross-linguistic comparison comes from Roark and Demuth (2000). The background for their study are the two generalizations given below:

98) *Two generalizations about cross-linguistic order of acquisition*

- a) Initial unstressed syllables appear in Spanish before Germanic (Lléo 1997,1998; Lléo and Demuth, 1999; Demuth, 2001)
- b) Coda consonants appear in Germanic before Spanish (Lléo et al, 1996; Lléo and Prinz, 1997; Lléo and Demuth, 1999)

In Spanish, initial weak syllables begin to surface somewhere between 1;6 and 1;10. For example, Lléo (2003) finds that roughly 40% of the utterances from her two Spanish-learning children at 1;6 already contain initial unstressed syllables. In contrast, English unstressed initial syllables appear somewhere shortly after 2. Trevor and Julia both begin to produce them at around 2;0; Gerken (1994) reports them appearing as late as 2 and a half. Meanwhile, Spanish codas are learned starting around 1;10 at the earliest; as seen in Lléo (2003)'s data in §2.3.1, José doesn't get them before 2;0 (see also Lléo 1997; Gennari and Demuth, 1997). In English, however, codas usually appear before or at the middle of the second year – for example, both Trevor and Julia acquire singleton codas between 1;4 and 1;6. It has also been reported for both German (Grijzenhout and Joppen, 1998) and English (Salidis and Johnson, 1997; Velleman and Vihman, 2000, 2002b) that some children's very first productions contain codas. In addition, Lléo et al

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

(1996) report that the proportion of closed syllables is already significantly higher in German than Spanish at the 25 word mark.⁴⁶

With these findings as background, Roark and Demuth (2000) compare data from Spanish and English to demonstrate that these structures are highly frequent in the language in which they appear first: initial weak syllables in Spanish, and codas in English. Their data came from corpora of child-directed speech in the two languages, using two lexicons of 18,000 words each extracted from CHILDES (thus admittedly reflecting token, not type, frequencies.)

First, they found that English initial syllables were stressed in an overwhelming portion of the corpus – monosyllabic words and disyllabic trochees already accounted for about 90 percent of the English tokens, (even when they allowed for the possibilities of encliticized 'the' and 'a'). In Spanish, however, 40% of the data contained initial unstressed syllables.⁴⁷ Furthermore, of the other 60% which had stressed initial syllables, more than a third of tokens (26% of the total) were due to just 10 extremely common words, mostly functional items: *con, en, es, no, por, que, sí, ver, y,* and *ya*. In contrast, 59.3% of the English words in that same sample had coda consonants, while only 25.2% of the Spanish words had codas.

A related study by Kirk and Demuth (2003) also found that the prevalent tendency of English children to learn complex codas before onsets (which was also true of Trevor) correlates with the frequency of these two structures in child-directed speech, rather than any of the other possible predictors they consider.

⁴⁶ Shelley Velleman (p.c.) reminds me that these differences also appear in babble.

⁴⁷ Of these 40% -- 10% were wS words, almost 20% were wSw, and the remaining 10% were longer words with initial w.

The overall finding here is that the frequency with which (at least some) marked structures occur in children's inputs correlates closely with their relative order of acquisition. In the Error-Selective Learning approach this connection is predicted, because markedness constraints that are violated frequently in the input will frequently create errors and so reach their Violation Thresholds earlier than infrequent ones.

A corollary of this connection is that structures with equal frequency in the child-directed lexicon should be variable in their order of acquisition. This prediction is discussed explicitly in the work and interpretation of Levelt and van der Vijver (2004), which discusses the order of acquisition of complex onsets vs. codas in the Fikkert/Levelt corpus. Among the 12 children acquiring Dutch in that sample, 3 acquire CCVC syllables before CVCC, while 9 acquire CVCC before CCVC. Since the frequency of both these syllable types is comparable in the child-directed Dutch they report on (3.4% of their Dutch corpus being CCVC vs. 3.7% being CVCC), the ESL explanation of this variation is the particular frequency quirks of the Error Cache that each child builds, based on their individual lexicons and experience.

4.2.2 Ambient not output frequencies, and the Error Cache

What we have seen above is that the lexical (token or type) frequencies in child-directed speech seem to be the driving factor. In the ESL theory of development, it is worth considering how those frequencies get mirrored in the errors that make it into the Cache.

The simplest assumption would be that children produce words with about the same frequencies as they hear them, and that the Error Cache is populated by all and only

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

children's overt production errors.⁴⁸ If the child's own production error frequencies were the right predictor, then if a child's lexicon were skewed – e.g. his words included a great many complex onsets and only a few complex codas – they would be predicted to acquire the former before the latter.

But looking at Trevor and Julia's outputs demonstrates for certain cases that *ambient* input frequencies are what drive the triggering process, and *not* a particular child's production error frequencies. The clear counter-example: Trevor uses more words with complex onsets than complex codas (and by token, MANY more – see table 97). Nevertheless, he acquires complex codas around 1;8-1;9, whereas complex onsets do not appear nearly at all in his productions until 2;2 (and not reliably until about 2;4), as in table 100).

99) *Complex syllable margins in the targets – Trevor*

age	<i>by word token (during the stage)</i>		age: up to...	<i>by word type (totals)</i>	
	CompOnset	CompCoda		CompOnset	CompCoda
0;10-1;3	101	15	1;3	11	7
1;4	51	17	1;4	13	8
1;5	94	41	1;5	24	15
1;6	59	54	1;6	28	20
1;7	96	90	1;7	38	32
1;8	116	82	1;8	53	45
total	517	299			

⁴⁸ Related to this line is the argument in Fikkert and Levelt (to appear) that the prevalence of certain places of articulation in a child's early lexicon dictates their patterns of consonant harmony.

100) *Complex syllable margins in the outputs - Trevor*

	Complex onset inputs			Complex coda inputs			
	C..	CC...	% CC	...0	...C	...CC	% CC
up to 1;3	101	0	0.0	11	3	1	6.7
1;5	93	1	1.1	10	25	5 ⁴⁹	12.5
1;6	59	0	0.0	4	29	14	29.8
1;7	91	2	2.2	12	46	29	33.3
1;8	104	11	9.6	6	29	45	56.3
1;9 ²	100	9	8.3	5	21	89	77.4
1;10	154	15	8.9	3	34	170	82.1
1;11	86	20	18.9	4	14	93	83.8
2;0	127	14	9.9	6	43	66	57.4
2;1	149	31	17.2	5	57	110	64.0
2;2	118 (19) ⁵⁰	30	20.3	2	49	112	68.7
2;3	87 (31)	46	34.6				
2;4	74 (25)	80	51.9				
2;5	31 (5)	78	71.6				

The most immediate consequence of this point touches on the nature of my proposed Error Cache, since it is the locus of frequency effects in ESL. What this data suggests is that we must understand the errors in the Cache to include not only overt production errors made by the learning child, but crucially also errors resulting from passive listening. One potential source of these errors might be the early application of a perception grammar of the sort proposed in Pater (2004) or Boersma (2001). The general idea is that perception errors are created when the child correctly hears a target word at some auditory level, but feeds that form as an input to their current perception grammar, and then notices that their perceptual grammar has mapped the phonetic form

⁴⁹ All 5 of these are productions of the word 'bump'.

⁵⁰ Of the total reduced clusters, the number in brackets are the reduction of 'tr' in 'Trevor', which around 2;2 he began to pronounce more than half the time with an initial [t] or [tʃ].

unfaithfully. From this noticing would come a silent ERC row – normal in all respects except its lack of phonetic implementation – which the learner will add to their Cache.⁵¹

Including such passive errors in the Cache also allows us to interpret the fact that at the very first stages of production, some children have clearly demoted some markedness constraints below conflicting faithfulness constraints (like NoCoda in English and German, as cited in section 4.2.1.⁵²) While children may never have produced words with codas before – they have surely heard many many words with them, and had time to build up enough errors to trigger learning on NoCoda. (Alternatively, the necessary errors might indeed have come from production, in late stages of canonical babble.)

However: this notion of including *perception* knowledge in the mechanism used for learning a *production* grammar clearly raises deeper questions about how these two kinds of phonological knowledge interact in development – questions that I take to be fundamentally unanswered (though see especially Pater 2004, Pater, Stager and Werker 2004, and also Escudero and Boersma 2003.) Research over the last decades has shown that children are sensitive enough to the frequencies in their linguistic input to prefer more frequent structures very early in life (see e.g. Jusczyk, Frederici et al (1993); Jusczyk, Luce and Charles-Luce (1994); Jusczyk (1997) and references therein.) Clearly, however, this sensitivity and awareness does not build the learner a fully-target grammar

⁵¹ One important consideration is that depending on the theory, the kinds of faithfulness constraints used in the perceptual parsing grammar may be different from those used in creating production errors – this is definitely the case in Boersma (2001) and to a lesser extent in Pater (2004) – so the kinds of ERCs rows produced in early perception will not necessarily mirror attested early errors in production. The spelling-out of this consideration and its consequences will be left unresolved here.

⁵² With respect to this early F >> M ranking, see also the different approach to stages of acquisition in Bernhardt and Stemberger (1998).

constraints are in a stringency relationship, but because of the frequency of violation of wordshapes.

At the initial stage, these learners reduces outputs to a single syllable. If we interpret this syllable as a bimoraic foot, this output satisfies a number of prosodic constraints: the need to align all feet with the word edge, and the demands of *both* foot form constraints, Trochee and Iamb (thanks to John McCarthy for suggesting this analysis):

103) *Prosodic Markedness constraints*

- a) All-Ft-L: “The left edge of every foot is aligned with the left edge of a Prosodic Word”
- b) Trochee: “Heads of feet must be left-aligned in the foot”
- c) Iamb: “Heads of feet must be right-aligned in the foot”⁵³

In the ranking below, Max ranks below all of these markedness pressures, so the winning candidate in the tableau of 103) is the single syllable output in (i). For illustration’s sake, I illustrate this with a three syllable word with medial stress, although the ranking generalizes to other multi-syllabic inputs:

104) Stage One: All-Ft-L, Trochee, Iamb >> Max

⁵³ There are clearly other differences between trochees and iambs than the alignment of their heads – see e.g. Hayes (1995). I assume these differences are the result of other constraints.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

105) *Stage one: one syllable only*

/σσσ/	All-Ft-L	Trochee	Iamb	Max ⁵⁴
(i) σ (σ _H)				**
(ii) (σ σ)		*!		*
(iii) (σ σ)			*!	*
(iv) σ (σ σ)	*!		*!	

106) *The resulting ERC row*

	All-Ft-L	Trochee	Iamb	Max
σ(σσ) ~ (σ _H)	L	e	L	W

At the intermediate stage, feet must still be aligned to both word edges (so there can still only be one of them). What characterizes the move to this stage, however, is that the language-specific foot type – either trochees or iambs – has been acquired, via demotion of the conflicting markedness constraint. In English, this intermediate stage will now treat the word-medial case above by producing a bi-syllabic trochee (tableau 108 below) – but it will still reduce any word with two feet to only one (109):

107) Stage Two: All-Ft-L, Trochee >> Max >> Iamb

108) *The intermediate stage: one trochaic foot...*⁵⁵

/σσσ/	All-Ft-L	Trochee	Max	Iamb
(i) (σ _H)			**!	
(ii) σ (σ σ)			*	*
(iii) (σ σ)		*!	*	
(v) σ (σ σ)	*!	(*)		(*)

⁵⁴ Note that I am calculating Max violations here in terms of syllables only because the candidates have been simplified to syllables. The real Max constraint I am assuming in fact counts violations by segment, but nothing crucial hinges on that here.

⁵⁵ The fact that outputs do not contain any post-tonic syllables at this stage, i.e. [(σ)σ], can be attributed to Parse-σ (Prince and Smolensky, 1993) or Lapse constraints (Elenbaas and Kager, 1999) constraints, that ultimately prefer to delete syllables if they cannot be footed.

109) ... and one foot only

/σσ'σσ/	All-Ft-L	Trochee	Max	Iamb
(i) (σ _{μμ})			**!	
(ii) σ (σ σ)			*	*
(iii) (σ σ)		*!	*	
(iv) (σσ)(σσ)	*!			*

The question is why children should decide to demote Iamb before All-Ft-L, and the Error-Selective answer comes from the frequencies of the forms that cause these two constraints to assign Ls. To see this, we must consider the kinds of errors in which these constraints have different violation profiles.

Since English is a fully trochaic language, Iamb is rampantly violated in the words that children hear. One very common English word shape is the trochee itself – bisyllabic, with initial stress. And these words will create ERC rows in which Iamb assigns an L but All-Ft-L does not, since both the winner and loser’s only foot is indeed left-aligned:

110) *The kind of English ERC row in which only Iamb assigns an L*

'máma'	All-Ft-L	Trochee	Iamb	Max
(σσ) ~ (σ _{μμ})	e	e	L	W

However, English is also a language with iterative footing, so that many winners violate All-Ft-L as well. But as we’ve seen above, the only English foot type that does not violate this definition of Iamb is a heavy monosyllable, so it will only be words with two monosyllabic feet to which only All-Ft-L will assign an L:

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

111) *The kind of English ERC in which only All-Ft-L assigns an L*

'pòntóon'	All-Ft-L	Trochee	Iamb	Max
(σ)(σ) ~ (σ _{μμ})	L	e	e	W

Since the Error Selective learner demotes constraints only once they reach the Violation Threshold, the question which type of ERC row will be more frequently represented in the Support? It is hopefully not contentious that children encounter English words like “mama” more frequently than words like “pontoon”. As we already saw in the Roark and Demuth (2000) findings of §4.2, 90% of their corpus of English child-directed speech contained tokens that were either monosyllabic (violating neither constraint) or disyllabic trochees (like the ERC row in 110). From this alone we should expect that Iamb will reach its Violation Threshold before All-Ft-L, and so add an error like 110) to the Support to create the attested intermediate stage.

To recall the more general point, then: stringency between markedness constraints is not in any way necessary for the ESL learner to pass through an intermediate stage of constraint ranking.

4.3.1 Noting an stringent alternative

It should be noted that Curtin and Zuraw (2001) in fact derive the one-foot intermediate stage using two markedness constraints on prosodic structure that sit in a stringency relation:

112) *The Specific M analysis of the one-foot stage* (from Curtin and Zuraw, 2001)
All-Ft-L >> Max >> All-σ-L

However, the less stringent constraint that they use is the somewhat implausible ‘All-Syllable-Left’, which requires every syllable to be aligned to the left edge of the Prosodic Word. The cross-linguistic support for this constraint is not particularly robust; those patterns that produce maximally one-syllable outputs may well be dealt with using the kind of prosodic constraints assumed in the analysis above (on prosodic maximality, see McCarthy and Prince, 1993; Ito, Kitigawa and Mester, 1996; Ussishkin, 2000; de Lacy, 2004.) I will return to Curtin and Zuraw’s analysis in chapter 3 §3, however, in the discussion of how stringency relations among faithfulness constraints shape the stages of GLA learning.

4.4 Infrequent mistakes and the value of the Error Cache

In the original BCD model, a one-time mistake in the data can in fact threaten the entire delicate search for restrictiveness. For example, if the BCD learner of a language with mid vowels only in stressed syllables happens to hear e.g. a slip of tongue with an unstressed mid vowel and add it to their Support, it will end up with the over-generating grammar that we have been trying so scrupulously to avoid.⁵⁶

We have already seen that adding a Cache to the BCD learning procedure means that not every error the learner makes will be learned from; in fact, many errors will be added to the then-current Error Cache, but never get transferred to the Support. So in the Error Cache, the learner also has a place to keep temporary track of the frequency of individual ERC rows – that is, how many times a particular winner-loser pair has been seen. Thus, one could include an initial criterion in the ESA saying that a best error (i.e. a

⁵⁶ As pointed out by Boersma and Hayes (2001), this ability to be ‘robust’ in the face of noisy data is a virtue of the Gradual Learning Algorithm – see chapter 3.

best error *type*) can only be one that has been made more than some minimum number of times (i.e. *tokens*.)

If we make this move, then to be make sure that the learner will still end up eventually with an empty Cache – see §3.2.4 above – we must also add some requirement ensuring that any ERC row which has been heard fewer than a certain number of times over a certain amount of time is erased from the Cache without any other impetus. In other words, hearing an insufficient number of tokens of a particular error type will lead the learner to decide that ERC row was just noise. Error-Selective learning makes this approach possible, unlike in BCD, because it decouples the reason for re-ranking, which is the current grammar’s errors, from the trigger of re-ranking, which is exceeding the Violation Threshold for some constraint. This gives the learner some leeway to ignore infrequently-made errors.

A very useful effect of keeping track of token as well as type frequencies in this way is that the Error-Selective learner can make the crucial distinction between (i) noisy data, which should never be transferred from Cache to Support, and (ii) grammatical exceptions, which should. Suppose that the learner is acquiring a language where a very few lexical items have codas – perhaps only three very recent borrowings – but 99.9% of the lexicon is coda-free. To be properly robust, the learning algorithm must be able to distinguish the Support for this exceptional coda language from one in which codas are 100% ruled out, even if the learner has misheard three words in the latter language as having codas. The difference will be found in their token frequencies. In the former language, only three ERCs can demonstrate the exceptionality of NoCoda but this exceptionality will be demonstrated *every* time each of these lexical items is heard,

whereas in the latter language the misheard codas will be one-time events. Thus as the former learner's VT sinks towards one, the three errors demonstrating the exceptional need for NoCoda >> Max will eventually trigger learning and get added to the Support, prompting some change to the grammar (recall chapter 1 §2.2.) In the latter case, however, by the time the VT gets low enough the Error Cache will already have been emptied of the misheard 'codas', just for having been heard only once.

5. Developmental variation and Error-Selective Learning

Perhaps the largest idealization made in the learning discussion of this chapter has been the abstraction away from any output variability in child data. The empirical reality is that children's outputs are in fact variable in a number of ways: that at any one stage of acquisition, children produce the same words or phonological structures in a variety of different ways.

As one example to use in the discussion that follows, I return to the first intermediate stage discussed in chapter 2 in which singleton codas have been acquired but complex ones have not. One of the children in table 3) of §2.2.1, P.J., was in fact at a stage where input singleton codas were only sometimes preserved, and other times deleted. Looking back at the data from Trevor and Julia's syllable margins month by month in section 3, it is clear that both children went through many months of variable singleton coda deletion. And after mastering faithfulness to singletons preservation, they later also passed through a stage of variability in their production of *complex* codas.

5.1 The ubiquity and challenges of variation in learning

The issue of where or how developmental variation should be captured by a grammatical learning theory does not seem straightforward. Broadly speaking, I see two ways into the problem. One is to attribute variability in development to the learning mechanism, and not to the grammars constructed by those mechanisms. The other way is to make variation an inherent property of the grammars per se: as we will see in detail in chapter 4, this is the nature of the stochastic OT approach and the associated Gradual Learning Algorithm (Boersma, 1997).

A third position worth considering is the possibility that all variation in learning is the result of performance problems. Under this view, learners whose grammar has just re-ranked so as to permit coda consonants must still learn to produce the necessary articulatory gestures associated with those codas. It seems reasonable that articulatory pressures are responsible for some of the variation that learners display, and I do not have any perfect arguments as to why they could not explain *all* variation. I note, however, that the connection between input frequencies and order of acquisition does not appear to hold in the case of marked structures that present clear articulatory problems. For example, the English interdental fricatives are notoriously difficult to produce, and while they are extremely frequent in English inputs they are quite late to be acquired. Thus, it might be possible to diagnose a kind of variation that is attributable to performance problems, and still find other evidence of grammatical variability left unaccounted for; I leave this tentative suggestion as a question for further research.

In some sense, the most extreme version of the performance problems view is to abandon the notion of children's outputs as involving phonology at all. The claim is that

the amount of attested variation indicates that constraint rankings are not responsible for any stages of production; this is at least the position of Hale and Reiss (1998). While this tack leaves us fewer things to explain, it does so at the expense of understanding several things. First, it does not give us any way of explaining the ways in which children's outputs are not *more* variable – that is, that they are stable and systematic, at all but perhaps the earliest stages (c.f. Ferguson and Farwell, 1975). A related, more specified analytic disappointment is that it writes off the observation that children's developing grammars can often mirror and innovate patterns found in the typology of natural languages – including those beyond the target – as an accident of flapping meat and phlegm. And third, a performance-only view can not explain why children's innovative patterns and errors can reflect sensitivities to abstract properties such as the notion of morphological basehood (see evidence of such innovations and discussion of this point in chapter 4 §7.2).⁵⁷

5.1.1 The potential for a variable BCD learner

How can developmental variation be treated in the present system? As I have already stressed (or perhaps conceded) this dissertation is no way an empirical study of variability in phonological learning. But since Error-Selective learning is an attempt to model more of the human acquisition process than pure BCD, we should at least ask to what extent variability across stages can be captured by this theory.

Given that the grammars my BCD algorithm learns do not contain any variation, my error-selective learner can only demonstrate variation through some elaboration of the learning procedure. This BCD algorithm builds what I will refer to as ordinal rankings –

⁵⁷ Thanks to Joe Pater for pointing out this argument to me.

that is, each constraint ranks above or below another -- e.g. C1 >> C2 -- but there is no sense in which C1 can be *more or less* ranked above C2. As I will discuss in some detail in chapter 3, other theories of learning assume an OT grammar in which constraints are ranked on a numerical scale, so that it IS possible for C1 to be ranked a lot or a little above C2 – this is true of the Gradual Learning Algorithm (Boersma, 1997 *et seq.*) We will see in chapter 3 that the possibility of one constraint outranking another one just a little bit is how the GLA learner naturally shows variation between its intermediate stages over the course of learning.

If we are committed to an ordinal OT grammar – which all extant versions of BCD such as the one I have adopted here certainly are – then our learner does not have any way to build *rankings* that encode any degrees of vacillation between intermediate stages, analogous to the GLA. Instead, however, we can consider how we could modify the error-selective BCD learner's methods in order to derive the effects of variation between rankings. In the rest of this section I will suggest two such possible methods: neither is presented as a definitive approach to variable Error-Selective Learning, but together they may provide future areas of investigation for the model.

The first alternative is to change the notion of a Violation Threshold from a fixed value to a range of values – this means that it will sometimes be easier to trigger learning and add new errors from the Cache into the Support than other times. If the learner temporarily adopted a low VT, they would add more errors to their Support and so build rankings that appear to represent a later stage of development. If at the same time the learner also remembered that the VT that allowed those errors into the Cache was lower than normal, they could periodically empty their Support of such suspect errors, and thus

build a ranking that reverts to an earlier stage. An initial implementation of this approach is given in 5.2 below; in section 5.2.4 I raise a few ways in which a more realistic version of this variable learner could be built.

A different idea about variation in ESL would be to suggest that the Support is not the single repository of permanent errors that I have been claiming it to be thus far. Instead, this variable learner would learn from a best error not by adding that error into the one Support but *cloning* the previous Support and adding the new error to that clone. In this approach, every cycle of learning would build a new Support (based on the previous one), BCD would be used to build a ranking tied to each Support clone, and learners could pick (randomly or otherwise) from their current ranking options in order to process new data. Over time, each Support would decay in memory as a function of how many errors it still made: the more errors, the quicker the memory loss. Once a Support was forgotten, its ranking would be forgotten, too, and so over time the older rankings would disappear from use and the newer ones would gain credence. This idea is briefly explored in section 5.3.

5.2 Alternative I: the Variable VT approach

As discussed in section 4, the introduction of an Error Cache has consequences for any aspect of learning that is in some way temporary. What I will explore here is the notion that errors in the Cache could derive the effects of later stage rankings by being temporarily introduced into the Support, but not retained because they have yet to truly overcome the Violation Threshold.

5.2.1 The example of variable codas

Recall the first Error Cache used to illustrate ESL of codas vs. complex codas, repeated below in 113). Up until now we have treated the Error Cache as inert up until the point when some constraint exceeds its Violation Threshold. So if, for example, our Violation Threshold is set to 4, then the Error Cache below is *about* to trigger learning on NoCoda but hasn't yet, and none of its errors have yet had any effect on re-ranking:

113) (repeated from 69)

<i>Input</i>	<i>Winner ~Loser</i>	NoCoda	<i>*CompCoda</i>	<i>Max</i>	<i>*CompOnset</i>
i) /frend/	frend ~ fe	L	L	W	L
ii) /piz/	piz ~ pi	L	e	W	e
iii) /gre'p/	gre'p ~ ge	L	L	W	L
iv) /ti/	ti ~ si	e	e	e	e

In the variable ESL approach, however, overcoming the true Violation Threshold is not actually necessary to trigger the inclusion of an error into the Support. Imagine instead that every time the learner uses the grammar they adopt a *temporary* Violation Threshold, that may be different than the true VTs (more on how this works in a minute.) If a temporary VT is lower than the true one, it may already be met or exceeded by some constraint in the Cache and thereby trigger an early application of the ESA. This early version of the ESA analyzes the Cache to find a best error – one which violates a Triggering Constraint according to the temporary VT. The learner will then *copy* this error (rather than *move* it as in normal ESL) to the Support.

In this system, a temporary trigger constraint is one whose number of Ls meets or exceeds the *temporary* VT; thus in 113) above, NoCoda is a temporary trigger constraint,

because its three Ls in the Cache meets the temporary VT. With this slight re-definition of triggering, the Early ESA is otherwise exactly the same as the original in §3.1:

114) *The Early ESL Algorithm*

Choose as the best error that row in the Cache which:

- a) has an L assigned by the *Temporary* Trigger Constraint and of those, the one that
- b) has the fewest Ls assigned by other Markedness constraints and of those, the one that
- c) has the most Ws assigned by Faithfulness constraints

As we saw when we were assuming a VT of 3: the Best Error in the Error Cache above is candidate (ii) “peas”, because it violates the temporary Trigger Constraint (NoCoda) but no other Markedness constraints. And in this Variable ESL scenario, “peas” is now the temporary Best Error.

115) *Using temporary violation thresholds to trigger Early ESA*

The true Violation Threshold: 4
The temporary Violation Threshold: 3

Temporary V.T
triggers learning

b) NoCoda is the Temp. Trigger Constraint

Input	Winner ~ Loser	No Coda	*Comp Coda	Max
/frend/	frend ~ fe	L	L	W
/piz/	piz ~ pi	L	E	W
/gre'p/	gre'p ~ ge	L	L	W
/ti/	ti ~ si	E	E	e

→
*early
ESA*

d) the Error Cache NOT cleared...

Input	Winner ~ Loser	No Coda	*Comp Coda	Max
/frend/	frend ~ fe	L	L	W
/piz/	piz ~ pi	L	e	W
/gre'p/	gre'p ~ ge	L	L	W
/ti/	ti ~ si	e	e	e

c) the pre-existing Learning Support Table

Input	Winner ~ Loser	No Coda	*Comp Coda	Max
... empty, waiting...				

e) ... but the Support IS updated:

Input	Winner ~ Loser	No Coda	*Comp Coda	Max
/piz/	piz ~ pi	L	e	W

Although this error’s Trigger Constraint may not yet have exceeded the true VT, and although the Cache has not yet been cleared – adding this Best Error to the Support will still trigger step 2, and so BCD will build a new ranking:

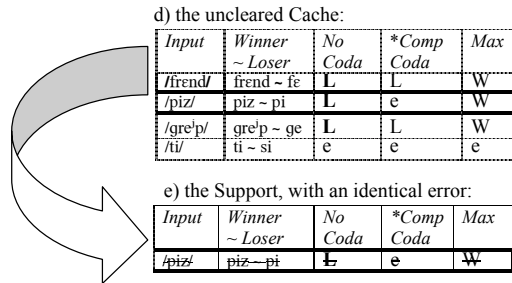
116) *The BCD ranking resulting from the new Support in 115)e):*

*ComplexCoda >> Max >> NoCoda

This new ranking in 116) has brought the learner to the intermediate stage of coda acquisition – but it is not a *stable* grammar because the learner has not yet seen enough errors to overcome the true VT and permanently demote NoCoda. In the Variable ESL system this instability has been encoded by copying rather moving temporary errors, leaving them in both the uncleared Cache AND the Support: while the learner relied on a temporary error to get their current ranking, they are not committed to its permanence in the Support. In the Variable ESL system, this instability is resolved at the end of each use of the current ranking using a process of ‘Synching’ the Error Cache and the Support. To perform this Synch, the learner compares the Error Cache and Support, finds any identical errors that appear in both, and *removes them from the Support*. This has the effect of forgetting any temporary errors that have been included in the Support due to a temporary Triggering Constraint.

In the example above: after building the ranking in 116), the learner synchs the Cache and Support, finds the identical ERC “peas” in both, and so deletes it from the Support:

117) *Synching the Cache and Support*



Now the Cache and the Support have been synched, the Support's evidence for Max >> NoCoda has been *removed*, so while the current ranking is at the intermediate stage of coda acquisition, the Support has returned to the previous stage. This means that the *next time* constraints get re-ranked (more on that below), the learner's ranking will follow the Support in reverting to the previous stage.

Where do temporary Violation Thresholds come from? Inspired by the stochastic OT system used by the GLA, where each run of EVAL randomly draws a value for each constraint from its probability distribution, I suggest that the Variable ESL begins each use of the grammar by similarly choosing a temporary VT value from a (normal) probability distribution, whose mean is the true Violation Threshold.⁵⁸

⁵⁸ If it turned out that different constraints should be assigned different Violation Thresholds, we could center this normal distribution around the *mean* of all the true VTs. For example, we might set the VTs for prosodic constraints lower than segmental ones, to derive the fact that learners acquire their stress systems much earlier than their full segmental inventories.

5.2.2 The effects of the Variable VT approach

I began this section with the idea that the Variable ESL learner chooses a temporary Violation Threshold every time the grammar is *used*, not only when it makes an error. So let us now step back and see how this system works as a whole, and then consider its pros and cons.

Every time the learner has used the grammar and chosen an optimal output for some input, they determine whether their output was an error or not. If they haven't made an error, they do nothing. If they have, they add the error to the Cache, generate a temporary VT, and then check whether any constraint in the Cache is now a trigger constraint. If not (that is, if no constraint has assigned as many or more Ls as the temporary VT value), then the learner simply goes to the existing, unrevised Support and builds a new ranking via BCD to be used next time. (This re-ranking gets the grammar back in synch with the Support, in case the last re-ranking was done using a previous temporary error.)

If the VT value chosen *has* been exceeded by some constraint or set of constraints, then the learner must add a best error to the Support. To know how to do so, the learner checks whether the temporary VT value is equal or greater than the true VT.⁵⁹ If it is, then the learner behaves as he or she would have in section 3: uses the normal ESA algorithm, *moves* the chosen best error into the Support, clears the Cache, and builds a new ranking via BCD to be used next time. If however the temporary VT value is *less* than the true VT, the learner instead uses the variable ESA algorithm, *copies* the best error into the Support, does NOT clear the Cache and again builds a new ranking via

⁵⁹ ... or the true VT for the relevant constraint, if assuming different ones.

BCD. Finally, the learner synchs the Cache and the Support so that a copied best error will be removed again from the Support.

The first thing to say about this proposal is that it has at least done what we set out to do. That is: adding a variable notion of the Violation Threshold, and synching the Cache and Support after learning, will indeed get us the effects of variation between stages. At the end of each learning cycle, the Support of a Variable ESL learner is in the same state that it would be under normal ESL. What's crucial is that its ranking may be different: if a low temporary VT led the learner to choose a temporary best error, the new ranking will reflect that error's ranking entailments but the Support will already have forgotten them (via synching). In the above example: the learner has built a grammar where singleton codas are preserved faithfully, but it has forgotten the error that enforced this faithfulness. For the moment the learner appears to have acquired singleton codas, and his grammar will parse them faithfully. But the next time the current grammar makes *any other* error a new ranking will be built from the Support, and the learner will return to a state of coda deletion. Through this flip-flopping of contents in the Support, this system derives variability between different rankings.

This extension of this model has also not sacrificed anything integral to the original ESL proposal. As in the original ESL proposal: when the *true* Violation Threshold is exceeded by some constraint, errors will permanently be moved into the Support, and all future rankings will reflect that move. Because the learner clears the Cache after applying the normal ESA, synching will not find any identical errors in the Cache and Support, so nothing will be deleted. And given that BCD will always choose a

ranking that prefers winners over losers, errors already in the Support will not be made again, so the Cache won't get re-cluttered with errors it has already (truly) learned from.

5.2.3 Deriving developmental regression in the variable VT approach

It might also be the case that the variable VT approach could be extended to explain apparent cases of regression, if we allowed the learner to adopt a low temporary VT *for an extended period of time* rather than choosing one after each new error. For example, if the learner above repeatedly chose a temporary VT of 3 for NoCoda they could appear to have fully progressed to the singleton coda stage – NoCoda would be continually triggering learning and adding errors with codas to the Support, and the synching progress would be removing those errors from the Support again, just as continually. If then the learner abandoned their temporary threshold and re-adopted the true, higher VT, NoCoda would stop triggering learning until the real VT was met, and in the meantime our child would appear to have regressed back to the coda-less initial stage.

What exactly would prompt the learner to adopt this lower VT for a long period of time, and why they would later revert to the true VT, remains unclear. As suggested in section 3.2.4 above, this proposal would be supported by evidence that children's regressions coincide with increased demands on their cognitive resources more generally – for example with the advent of a burst in lexical acquisition.⁶⁰

⁶⁰ See the somewhat related arguments in Stager and Werker (1997) and Fennell and Werker (2003) about the connection between decreased phonemic discrimination in tasks that pair sound and meaning among infants who have reached a stage of increased lexical acquisition (around 14 months). See also Pater, Stager and Werker (2004) for discussion of OT implementation of the relationship between cognitive load and variable rankings.

5.2.4 Weaknesses of the variable VT approach

This section has sketched one direction in which a variable ESL learner might evolve, but this approach does not satisfactorily address all the issues. For one thing, the use of variable VTs is somewhat stipulative: especially because if the learner can always reference what the real VT is and use it to re-synch the Cache and Support, it is somewhat unclear why they would periodically choose a temporary lower one. Furthermore, although picking the temporary VT from a normal distribution predicts that most temporary values chosen will be clustered around the true VT value, this approach doesn't really connect the degree of evidence the learner has for ranking with the likelihood that they use that ranking at any given time.

5.3 Alternative II: the Cloned Support approach⁶¹

In addition to the variable VT idea, section 5.1.1 also raised a second possibility about developmental variation in BCD learning. This alternative retains the original ESL ideas of a single, true Violation Threshold to trigger learning, and a single mechanism by which chosen errors are added permanently to a Support and the Cache cleared. What is different in this account is the conception of the Support itself. This learner uses each new best error to to build an alternative Support, which contains all the old errors plus the new one, and which is kept in memory alongside the previous one. Each Support is used by BCD to build a grammar, and each time the learner goes to produce a new output they can choose any of the currently-held grammars to feed it through. Thus in this model, the learner varies between stages because they vary their choice of stored grammar to use.

⁶¹ Thanks to Lyn Frazier and Jonah Katz for comments that inspired this approach.

5.3.1 Returning to the variable coda example

To see how the cloned Support approach works: suppose that our learner's Violation Threshold is 4, and that the learner has just added an error to their Cache that will trigger learning on NoCoda

118) *An Error Cache in which NoCoda overcomes the VT:*

Input	Winner ~Loser	NoCoda	*CompCoda	Max	*CompOnset
i) /frend/	frend ~ fe:	L	L	W	L
ii) /piz/	piz ~ pi	L	e	W	e
iii) /gre'p/	gre'p ~ ge	L	L	W	L
iv) /tost/	tost ~ to	L	L	W	e

Looking at these errors, we can see that while the learner has yet to learn much of anything about English syllable structure, it has already acquired some simple facts about the English segmental inventory – for example, that mid vowels and labial consonants are all allowed. This means that some errors demonstrating a tolerance for these marked features must have already made it into the Support (as in 119a below), building a grammar like in 119b):

119)a) *An existing Support for the learner in 118)*

Winner ~ Loser	*Mid	*Lab	Ident [mid]	Ident [lab]	No Coda	*Comp Coda	Max
bé'bi ~ d'idi	L	L	W	W	e	e	e

119)b) *A grammar that BCD builds from 119a)*⁶²

NoCoda, >> Id[lab] >> Id[dors] >> *Mid, >> *Max
 CompCoda *Lab

As discussed several times already: the ESL learner faced with the Cache in 119) will choose ‘piz ~ pi’ as the best error to learn from, and up until now that has meant updating the Support with this error. Instead, this alternative learner uses the best error from 118) to build a clone of the Support in 119)a), and build another ranking from that clone. This means that after NoCoda overcomes the violation threshold in 118) and a cycle of learning has occurred, the learner has TWO Supports, as in 120) below, and thus that its grammar contains TWO different rankings as in 121):

120) *The state of the cloned Support learner after NoCoda triggers learning in 118)*

a) *Support A – pre-existing*

Winner ~ Loser	*Mid	*Lab	Ident [mid]	Ident [lab]	No Coda	*Comp Coda	Max
bé'bi ~ didi	L	L	W	W	e	e	e

b) *Support B – cloned Support A plus one new error*

Winner ~ Loser	*Mid	*Lab	Ident [mid]	Ident [lab]	No Coda	*Comp Coda	Max
bé'bi ~ didi	L	L	W	W	e	e	e
piz ~ pi	e	e	e	e	L	e	W

⁶² A reminder of how BCD gets a grammar like this from 117a). First we install all M constraints with no Ls (those against syllable structure); then we have to install one F constraint that assigns a W, so we install Id[dors] to free up *Dors in the next stratum. Then we again have to install F constraints until an M constraint is available, which means Id[mid],[lab] to free up *Mid, *Lab, and then we install the remaining F constraint Max and we have a grammar.

121) *The resulting grammar with two rankings*

a) from Support A: **NoCoda**, >> Id[mid] >> Id[lab] >> *Mid >> **Max**
 CompCoda *Lab

b) from Support B: CompCoda >> **Max** >> **NoCoda** >> Id[dors] >> Id[lab] >> *Mid
 *Lab

The bold face and underlined constraints are those in conflict with each other. This is to make clear that the first ranking in 121a) is one where segmental restrictions have been overcome (F >> M) but syllable structure remains fully unmarked (M >> F) – while in 121b) some syllable markedness has also been acquired (specifically Max >> NoCoda).

In this ESL model, the learner now has two rankings as part of their grammar, and every time it uses its grammar it must first pick one of its rankings. When it picks the one built from Support B, it produces singleton codas faithfully; when it picks the one built from Support A, it still deletes all codas. And thus it vacillates between two intermediate stages.

Note that to remain in line with the goals of this dissertation, our learner must still be remembering Support(s) as its primary data rather than rankings – so, we can say that though the learner has multiple Supports in memory simultaneously, it also knows which ranking comes from which Support, and as soon as any Support is forgotten its associated ranking disappears as well.

Thus, the necessary second part of this cloned Support model is how the learner gets rid of old Supports. The basic proposal is that each Support decays in memory in proportion to how many errors it prompts the learner to make. One way to implement this idea would be that the learner keeps a “reliability score” associated with each current

Support hypothesis⁶³ – suppose we start each freshly-cloned Support’s score at 1 (meaning 100% reliability). The first time a particular Support’s ranking is used to process an observed form and makes an error to add to the Cache, that Support’s reliability score is lowered: perhaps by a fixed amount (say to 0.9), or perhaps more intelligently as a proportion of the number of errors in that Support.

In this second scenario: suppose the freshly-cloned Support B in 120) had a reliability score of 1 and we then used Support B’s ranking to make a new error like [tost] ~ [tos]. The learner would now have the Support’s two resolved errors in favour of the ranking (on ‘baby’), and one new error against it (on ‘toast’), so its reliability score would now be 0.5 (1 out of 2).

Finally: once a Support’s reliability score sinks low enough it is forgotten altogether, and its associated ranking disappears as well. In the case of 120), the ranking built from Support A makes all the same errors as that built from Support B – *plus* errors on singleton codas. Thus Support A’s reliability score will sink faster than Support B’s. Once Support A is forgotten, the learner will have moved out of the vacillation stage, and always produces singleton codas faithfully from now on.

5.3.2 Discussion of the Cloned Support approach

This alternative provides a different view of variation in ESL than the variable VT approach. This most recent learner does not vary between stages because the contents of their single Support grows and shrinks again, but their *set* of Supports grows and shrinks. One benefit of the cloned Support approach is that it does not require any selective amnesia of the true VT; nor does it require any process like synching.

⁶³ The idea of a reliability score comes quite directly from Albright and Hayes (2003)’s rule-based learner.

In this Support-cloning view, all variation is dictated by the order and speed with which new Supports are created and old Supports are forgotten. New Supports are still built in the normal ESL fashion: i.e. when learning is triggered by some constraint overcoming the VT in the Cache. Meanwhile, old Supports are forgotten via their reliability score. This ensures that older Supports – ones that have fewer target rankings and so prompt more errors – are forgotten quickly and newer Supports are retained longer. In the end, the learner’s final Support will retain a perfect reliability score – because it never makes any new errors.

5.3.3 Regression in the Cloned Support approach

Another use of the reliability score could be to influence the learner’s choice between its multiple current Supports in processing new data – the higher a reliability score, the more likely the learner could be to use that Support’s ranking. A side benefit of connecting a Support’s reliability with its ranking’s continued use might be that quirks in the data could create regressions to earlier stages.

To get regression in the cloned Supports model, the learner would have to get hung up on using an older Support rather than a newer one for a period of time. This would mean that an older Support would need a higher reliability score than its competitors, which could happen temporarily as a fluke of randomization. Suppose that two new Supports have just been created, so that each has a near-perfect reliability score and each associated ranking is being chosen about as often as the other. If it happened that most of the observed forms fed to the slightly older Support were relatively unmarked, while most of the marked forms were fed to the slightly newer Support, the

misleading upshot would be that the newer Support was *less* reliable. And for a short while – until the errors of the older Support caught up – the learner could appear to have regressed to an earlier stage.

5.4 Summarizing the variable ESL discussion

This section has presented some issues and ideas for extending the Error-Selective learner to model developmental variation between stages. I suggested two different ways in which the general proposal could be modified, by either adding some errors to the Support in a temporary way (§5.2), or by building multiple, temporary Supports (§5.3)

One point about both suggestions is that these variable ESL learners clearly treat variability between stages of acquisition differently than variation at the end-state grammar. Once errors are no longer being made, there will be no more errors in the Cache to violate Violation Thresholds or trigger Support clonings – so there will be no vacillation between rankings. This contrasts sharply with the GLA approach to be discussed in the next chapter, in which variation between rankings is an inherent property of every grammar: developing, stable or otherwise. The extent to which the variability seen in developing vs. adult grammars should be treated as a unified phenomenon is not necessarily clear – in part because adult speakers can overtly control their choice of variants with respect to socio-linguistic factors, in a way that a child varying between the codaless and singleton coda grammars clearly does not. Still the BCD learner's treatment of any kind of variation remain tenuous enough to require further work; after my discussion of the GLA, I return to the issues of end-state variation and BCD-style learning in chapter 3 §5.5.

6. Chapter Summary

The goal of this chapter has been to introduce Error-Selective Learning, as a framework for gradual learning using BCD. I have discussed at length the ways in which ESL uses properties of ERC rows and their frequency to slowly add errors to the Support, which in turn slowly provides evidence to the learner of the target grammar. I have exemplified the approach and the stages it provides using a number of constraints and languages from the literature, which I hope will have demonstrated its breadth. I have also introduced two alternative ideas for how the Error-Selective Learner could vary between stages in a gradual way, and even show the temporary effects of developmental regression. The best way to incorporate variation into the ESL model, particularly with the BCD's view of constraint rankings, still remains to be seen; see also chapter 3 §6.

CHAPTER IV

THE GRADUAL LEARNING ALGORITHM ALTERNATIVE

1. An introduction to the Gradual Learning Algorithm

Part of the argument put forward in the previous chapters is that an OT learner can learn partially from the set of available data, and in a frequency-sensitive way, while still using a classic OT grammar that does not encode frequency itself. In this chapter, I discuss an alternative approach in the OT literature that takes a very different view. This method is called the Gradual Learning Algorithm or GLA (Boersma, 1997; Boersma and Hayes, 2001; Curtin and Zuraw, 2001; Levelt and van der Vijver, 2004; Hayes and Londe, 2006). The GLA is fundamentally different than BCD – both in the kind of grammar that it learns, and the way it processes errors – but it has properties that make it very relevant to the issues discussed here so far.

In this section I will introduce the kind of grammar that the GLA learns – a numerical and stochastic brand of OT – and then the GLA algorithm itself. I will highlight how the GLA is inherently a stage-like learner, and consider the role of ranking biases in its workings. Note that this section is not intended to provide a comprehensive introduction to the GLA: see Boersma (1998); Boersma and Hayes (2001).

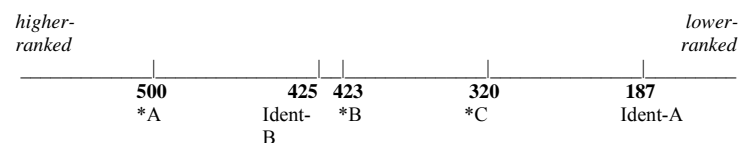
1.1 The GLA view of constraint rankings

The constraint rankings that the Error-Selective BCD learner learns are classic OT hierarchies in that its constraint rankings are *ordinal*. Two constraints in the true, classic OT of Prince and Smolensky (1993) can only stand in one of two relations – $A \gg B$ or $B \gg A$ – and there is no sense in which A can be ranked *more or less* above B or vice

versa. We have also seen that in the T/S view of learning, this property is relaxed slightly to allow a third relation: one of equal ranking. Thus, the first constraint ranking learned by the BCD algorithm is of the form $\{M\} \gg \{F\}$. In this hierarchy: for each M and F constraint pair, $M \gg F$, and within the M and F strata, each constraint is ranked equally with all others.

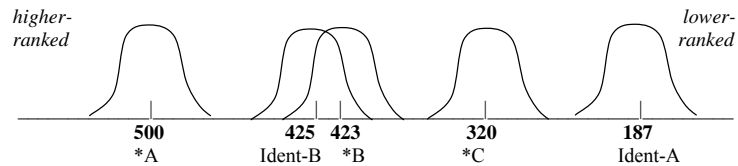
In contrast, the constraint rankings that the GLA learns are what I will call *numerical* rather than ordinal. The GLA learns grammars where constraints have ranking values along a number line, so that every constraint is ranked not just above or below every other, but *at a certain distance* above or below every other. This is shown below for some hypothetical constraints and ranking values:

1) The numerical view of OT constraint ranking (first try)



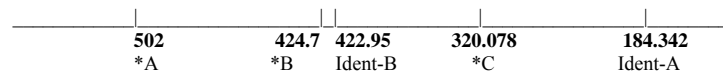
Furthermore, the GLA learner also assumes that constraint rankings are *stochastic* – that they are perturbed by some statistical noise. This noise is introduced by assuming that a constraint's ranking value does not just represent its single point on the scale, as in 1), but rather the midpoint of a normal (Gaussian) distribution of values. This means that constraint X's ranking value is the place in the hierarchy that X is the most likely to sit, and the further away from X's ranking value you get, the less likely X is to have that value at any point in time:

2) *The stochastic view of OT constraint rankings¹*



Each time a stochastic OT grammar is used, a single value is chosen for each constraint from its distribution of values; this choice creates a scale of single-point constraint values as in 3) below, which for practical purposes can be used by EVAL as a classic OT ranking:

3) *A one-time ranking*



4) *The ordinal version of 3)*
*A >> *B >> Ident-B >> *C >> Ident-A

(For reasons of typographic ease the GLA numerical rankings I draw from now on will not contain the normal distribution curves above each ranking value, but the GLA model I will be discussing throughout does indeed include this stochastic component – crucially so when treating variation in section 6.)

Despite the fact that each run of this grammar relies on a single ranking that can be equated with a classic OT hierarchy as in 4), there are crucial differences between the ordinal and numerical OT versions of OT. The example above has already demonstrated

¹ Note that the curves I have been able to draw freehand here are really not in the shape of a normal distribution at all.

this, with respect to the ranking of *B and Ident-B. The ranking values in 1) showed us that in this grammar Ident-B is ranked above *B, but only slightly above; this means that in 2), their distribution of values overlap considerably. Because of the stochastic component of this model, these similar ranking values mean that Ident-B is only slightly more likely to outrank *B in any run of the grammar: in the one-time ranking in 3), for example, the value chosen from *B's distribution is in fact *higher* than the one chosen from Ident-B's, so that for this use of the grammar, their ranking has been reversed. (Note that the amount to which the curves of two constraints overlap is a function not only of how similar their ranking values are but also how much random noise the system uses to choose one-time values.)

It is in this way that the relative distance between constraints makes numerical, stochastic OT different from the classic theory of Prince and Smolensky (1993/2004). It is also the conception of ranking values as numbers on a line that makes the Gradual part of the GLA possible, as we will now see in the next section.

1.2 How the GLA learns a grammar

Like BCD, the GLA is an error-driven online-learner: it notices when its current grammar produces a loser form, different from the ambient winner, and reacts to such an error by re-ranking constraints. However, the GLA's procedure of re-ranking is very different from the BCD one. Rather than using the Ws and Ls of winner-loser pairs as a starting point to find a new ranking, the GLA merely promotes *all* constraints that assign a W and demotes *all* constraints that assign an L.²

²This particular method of choosing constraints to promote and demote is really only one of many GLAs considered in Boersma (1997) and Boersma and Hayes (2001). However, this is the one that these authors

To illustrate: suppose a GLA learner whose current ranking values are those in 2) encounters the target form [A], and feeds /A/ to EVAL using the one-time ranking as in 3). This ranking will create the error in 5) below:

5)a) *An error caused by one run of the grammar in 3):*

winner~ loser	*A	*B	Ident-B	*C	Ident-A
A ~ C	L			W	W

In response, the GLA will now adjust ranking values accordingly; this process is called a learning trial:

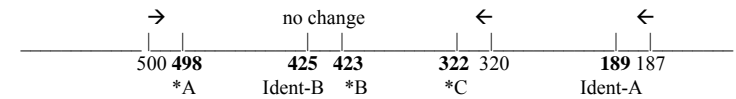
5)b) *The GLA's response to the error*

winner~ loser	*A	*B	Ident-B	*C	Ident-A
A ~ C	L →	(no change)	(no change)	← W	← W

How much are constraints promoted and demoted? Unlike in BCD, this is a question that must be answered, because our constraints are ranked by absolute values, not just relative to each other. The amount by which each constraint is moved in response to an error is referred to as the learner's *plasticity*, and the GLA assumption is that over time the learner's plasticity decreases, so that constraints move less and less in response to errors. If for example the learner's plasticity is currently 2, the actual re-ranking effect of 5)a) applied to the old grammar from 2) will be as in 6) below – here, the previous ranking values are in regular font, and the new values are in bold:

find works best – in particular, Boersma and Hayes (2001) diagnose this brand of GLA as the only one that produces the variation patterns they attempt to model -- and it's also the default version of the GLA used in OTSoft (see later this section). Therefore, I will refer to this re-ranking algorithm as "the GLA" from here onwards.

6) The new GLA grammar:



1.2.1 The (limited) power of an error in the GLA

The GLA learner does not attempt to *resolve* errors in any immediate way: the grammar in 6) is only very slightly less likely to make the error in 5a) as the previous grammar in 2) was. To remember how different this is from the BCD approach I adopted in previous chapters: suppose we had added the error in 5a) to our BCD Support instead:

7) *The initial ranking from 4):*

*A >> *B >> Ident-B >> *C >> Ident-A

8) *The error from 5a)*

winner~ loser	*A	*B	Ident-B	*C	Ident-A
A ~ C	L			W	W

9) *BCD learning result:*

*B, *C >> *A >> Ident-A, Ident-B

In the BCD approach, this error has been enough to *completely* re-arrange the grammar (compare 7) to 9), whereas the GLA learner has only slightly revised its ranking values.

Recall, however, that the BCD learner is not doomed if this re-ranking is wrong: since this error is stored in the Support, later re-rankings can undo any of these rankings if necessary. Not so in the GLA: the GLA learner does *not* store its errors for any later

use. After the learner has gotten to the ranking values in 6), it erases 5a) from its memory and starts creating new errors with its new ranking.

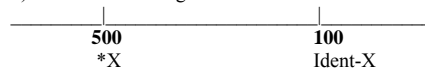
Over time, constraints that frequently assign Ls will move towards the bottom of the hierarchy and those that frequently assign Ws will move towards the top. In this way, the GLA learner demonstrates both intermediate stages and fluid grammatical variation. If the current grammar consistently produces errors where markedness assigns an L and faithfulness assigns a W, the ranking values for M and F will approach each other, cross over, and finally move away from each other. As a result, the learner's outputs will gradually shift from the M >> F grammar to the F >> M grammar, with variation between the two along the way:

10) *Demonstrating gradual learning in the GLA*

a) An error:

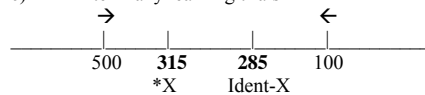
winner ~ loser	*X	Ident-X
X ~ Y	L	W

b) Initial ranking values:



The grammar's output
 /X/ → [X] almost never
 /X/ → [Y] almost always
 Classic OT analog: **X >> Y**

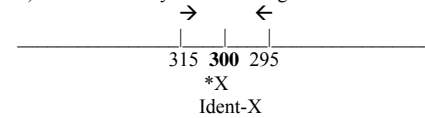
c) After many learning trials



The grammar's output
 /X/ → [X] occasionally
 /X/ → [Y] usually
 Classic OT analog: **variation**
 X >> Y, Y >> X

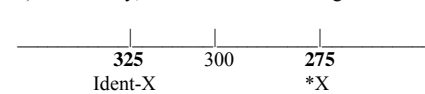
Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
 Ph.D. dissertation, UMass Amherst

d) After many more learning trials



The grammar's output
 /X/ → [X] half the time
 /X/ → [Y] the other half of the time
 Classic OT analog: **variation**
 X >> Y, Y >> X

e) Finally, after even more learning trials



The grammar's output
 /X/ → [X] almost always
 /X/ → [Y] almost never
 Classic OT analog: **Y >> X**

And now that /X/ is being faithfully mapped to itself almost always, the grammar is (almost) not making errors anymore (practically speaking)³, so learning is no longer being triggered, and we have reached the final grammar.

As the example in 10) has shown, the only way to see how the GLA learns is to give it lots and lots of learning trials and track its progress over time. This is best done using a computer simulation; in this chapter I will use the simulation in OTSoft 2.1 (Hayes, Tesar and Zuraw, 2003.)⁴ OTSoft takes an initial set of ranking values⁵ and a table of input learning data that essentially represent ERC rows (i.e. winners, losers, and constraint violations), feeds the GLA a specified number of learning trials picked from the ERC rows, and then returns a set of new ranking values. As we will see below,

³ For one constraint to outrank another in this stochastic OT model, practically speaking, means that their ranking values are far enough apart that choosing a ranking where their values are reversed is very, very, very unlikely. It is however true that normal distributions are asymptotic to zero: so in principle no set of ranking values makes any constraint reversal truly impossible, just very unlikely. This technicality will not concern us here.

⁴ The other major GLA software comes with Praat (Boersma and Weenink, 2006).

⁵ More on the initial ranking values I used below.

choosing different numbers of learning trials allows a window into the various stages that the GLA passes through along the way from initial to final states.

1.3 Goals and core properties of the GLA

The GLA was designed to be inherently stage-like in its learning. From the present perspective, the way the GLA achieves these stages also seems very reasonable: chapter 2 section 4 provided some compelling evidence that frequency of markedness violation is a good predictor of order of acquisition, and the GLA's method of demoting L-preferring constraints precisely encodes this correlation (more on this connection in section 2 below.) Relatedly, the GLA's use of stochastic constraint rankings allows the model to learn grammars with variation, in intermediate stages of acquisition and the end state. The GLA is also designed to be robust in the face of misleading learning data, such as slips of the tongue overheard by the learner, because (unlike classic BCD) no particular error makes much of a difference (though c.f. the ESL proposal in chapter 3 section 4.4.)

Two central properties of the GLA should be kept in mind throughout the rest of this chapter. First: the GLA is *inherently* about numerical rather than ordinal OT; the stochastic view of constraints is required to make “move M1 down *a little bit*” a coherent notion. Second: a fundamental difference between GLA and BCD learning is that the GLA doesn't save its errors. In the BCD view, the Support is a record of why the current grammar has deviated from its ranking biases in the way that it has, while rankings come and go. In the contrasting GLA view, the ranking values are themselves the repository of learned information; rankings change slowly, and they require the evidence of many

errors to be reversed. The hope is that this gradualness will be enough to ensure that incorrect re-rankings are prevented.

1.4 Ranking Biases and the GLA

How can the GLA incorporate ranking biases? In one sense it is easy. Any strictly initial state bias, like the Smolensky (1996) M >> F bias, can be mimicked in this system just by giving different initial ranking values to classes of constraints: setting e.g. all Markedness constraints to the initial ranking value 500 and all Faithfulness constraints to 100.

With respect to the GLA's need for this bias: Curtin and Zuraw (2001) point out that their learner eventually finds the final grammar without the M >> F bias (i.e. with constraints all beginning with the identical ranking value), but that its early stages fluctuate wildly in ways that don't match the attested data they are trying to model. And eventual convergence is certainly not guaranteed with just *any* distribution of initial ranking values for markedness and faithfulness. If in the target grammar M1 >> F1, but at the initial state they are ranked F1 >> M1, the learner will not make an errors showing that they need re-ranking – as we have seen many times, high-ranking F means no errors in phonotactic learning. So these two constraints will only get re-ranked appropriately if they conflict with *other constraints* that cause errors where M1 assigns a W.⁶

What the next two sections demonstrate is that in addition, the bias for F-context subsets from chapter 1 is also necessary for the GLA to reach the target grammar.⁷

⁶Since, as we've seen, IO-faithfulness can assign no Ls in phonotactic learning.

⁷So far as I know, only Hayes and Londe (2006) have made this point – I return to their approach using *a priori* rankings in sections 2.5 and 6.2 below.

1.5 Chapter Roadmap

Sections 2 and 3 focus on ranking biases in the GLA. Section 2 deals with end-state grammars and the bias' role in ensuring restrictiveness; section 3 returns to the Specific-F type of intermediate stages and demonstrates the GLA's difficulty in predicting such stages without the bias. In section 4 I discuss the ways in which persistent and/or contingent ranking biases might be imposed in the GLA, and the difficulties that these approaches face. In section 5, I return to the GLA's lack of stored errors, and the problems for restrictiveness and convergence that the GLA's brand of memory-less learning causes. In section 6, I compare and contrast the GLA and BCD learners with respect to two problems in later phonological learning: exceptionality and variation in end-state grammars. As with the previous ones, this chapter's discussion will highlight the learner's need for error memory, in a format like the Support.

2. Restrictiveness and specific-to-general faithfulness relations in the GLA

I illustrate in this section the fairly simple point that the GLA's reliance on frequency of violation makes it learn superset grammars as soon as specific faithfulness constraints are admitted into CON. To do so, I report a GLA simulation using hypothetical data, stopping every few 100 trials to see the learner's progress until an stable end-state grammar has been reached. The results also show how the GLA puts faithfulness constraints on the same stringency scale into general >> specific rankings – although whether this result is a problem for the end-state grammar is not clear.

2.1. The exemplifying grammar

The hypothetical language I will use in this illustration, imaginatively called L, has two sets of vowels that are assumed to be marked: mid vowels and front rounded vowels. In L, the mid vowels [e, o] are restricted to stressed syllables only, while the front round vowels [y, ø] occur in both stressed and unstressed contexts. Both of these properties are well-attested: the former in Southern dialects of Italian (Maiden, 1995; Flemming, 2001) and Russian (Halle, 1959, Flemming, 2001); the latter in e.g. French and Turkish.

To let us build some words and make errors: I will assume that the entire vowel inventory is a 7 vowel system of the form: [i, y, u, e, œ, o, a], and that stress is always word-initial. According to the two prohibitions above: mid vowels [e, o, œ] only appear in stressed environments, and the front rounded vowel [y] appears both stressed and unstressed.

Given these parameters, the possible words that our learner must therefore learn to produce are as in 11):

- 11) *Possible words of L, wrt mid/round vowels*
- | | | |
|-----|----------------|-----------------------------------|
| (a) | [képa], [kópi] | (stressed mid vowel) |
| (b) | [pœ́ki] | (stressed mid, front/round vowel) |
| (c) | [lýpi] | (stressed front/round vowel) |
| (d) | [pítý] | (unstressed front/round vowel) |

In order to learn a restrictive grammar, what we want our learners to realize is this other fact:

- 12) *Impossible words of L*
(a) *[kipe], *[pákœ] (no unstressed mid vowels)

To get this grammar, we first need general faith to vowel frontness and rounding to rank above *front/round, as in (13)a. To get the distribution of mid vowels, we need *[mid] to rank below the positional faith constraint, Ident-mid(σ'), but above general faith to [mid], as per (13)b):

- 13) *The necessary rankings of L:*
(a) Ident-rd(Seg), Ident-back(Seg) >> *front/round
(b) Ident-mid(σ') >> *mid >> Ident-mid(Seg)

Concentrating on just this portion of the grammar: the two driving forces in learning will of course be the two markedness constraints *front/round and *mid. To conflict with these markedness constraints, I will assume four Ident families of constraints – Id-Mid, Round, Front and Back – each with both a general and stressed-syllable version.

2.2 The GLA's learning input

To get started, we want to understand the errors that our learner will be making and learning from. Additionally, we need to see how the constraints that the GLA will be assuming work in treating the errors' winners. I assume that our GLA learner will be given a M >> F bias, so that *mid and *front/rd begin with ranking values of 500 and all Ident constraints at 100.

The four tableaux below present the initial errors that our learner will make. Two notes: first, I have included all seven vowel candidates in each case only to show clearly what I gave the GLA to teach it the appropriate constraint violations – candidates with equal or greater markedness than the inputs are shaded out to slightly illuminate the interesting candidates. Second: the first three tableaux deal with stressed vowels – in these cases, Ident(Seg) and Ident-σ' get the same violations, so I have collapsed them into one. In each case, the two stars Ident receives should be understood as one violation of the specific constraint, and one of the general.

As it turns out: given this constraint set, all marked vowels will map in the initial grammar to high unrounded ones (either [i] or [u], depending on the input value for front/back.) We will see this for each vowel in turn.

First: /'kepa/ has a stressed mid vowel; the best way to repair with this constraint set is to raise it to the high vowel [i], as all other vowels are more unfaithful (i, iii, iv) or more marked (v-vii):

14) *The tableau of violations for /'kepa/*

/kepa/	*mid	*front/rd	Id(mid)	Id(bk)	Id(rd)
(i) 'kepa	*!				
(ii) ☞ 'kipa			**		
(iii) 'kapa			**	**!	
(iv) 'kupa			**	**!	**
(v) 'kypa		*!	**		**
(vi) 'kœpa	*!	*!			**
(vii) 'kopa	*!			**	**!

(Note that given that all the faithfulness constraints being discussed here are symmetric: forms with the back stressed mid vowel [o] get the exact same treatment, raising to u).

The form /pæki/ has a stressed vowel that is both mid and also front-rounded; among the vowels that do not violate either markedness constraint (ii-iv), the best repairs are either of the high vowels, [i] or [u]:

15) *The tableau for /pæki/*

/pæki/	*mid	*front/rd	Id(mid)	Id(bk)	Id(rd)
(i) 'pæki	*!	*!			
(ii) \mathcal{E} 'puki			**	**	
(iii) 'paki			**	**	*!
(iv) \mathcal{E} 'piki			**		**
(v) 'pyki		*!	**		
(vi) 'peki	*!				**
(viii) 'poki	*!			**	

The form /'lypi/ has a stressed front-rounded vowel; again, among the unmarked vowels the best repair is unrounded [i] or rounded [u]:

16) *The tableau for /lypi/*

/lypi/	*mid	*front/rd	Id(mid)	Id(bk)	Id(rd)
(i) 'lypi		*!			
(ii) \mathcal{E} 'lipi					**
(iii) \mathcal{E} 'lupi				**	
(iv) 'lapi				**	*!
(v) 'lepi	*!		**		**
(vi) 'lopi	*!		**	**	
(vii) 'læpi	*!	*!	**		

Finally, the last form /'pity/ has a round vowel in an unstressed context, meaning that only the general Ident faith constraints are relevant to controlling its repair. Other than that, however, its violations are the same as for the /y/ in the stressed context:

17) *The tableau for /pity/*

/pity/	*mid	*front/rd	Id(mid) Seg	Id(bk) Seg	Id(rd) Seg
(i) 'pity		*!			
(ii) \mathcal{E} 'piti					*
(iii) \mathcal{E} 'pitu				*	
(iv) 'pita				*	*!
(v) 'pite	*!		*		*
(vi) 'pito	*!		*	*	
(vii) 'pitæ	*!	*!	*		

As with the BCD: the goal of the GLA is get from this state to one where all inputs map to the winners (rather than to these losers or any subsequent ones.)

2.3 **The stages of GLA learning**

2.3.1 **The initial stage**

The GLA was given an initial M >> F bias of 500 >> 100, so the initial ranking values for all of our mini-grammar's constraints were as in 18) below. Note that I have listed the constraints not according to their initial ranking but instead alongside the constraints they are in conflict with:

18) *Initial ranking values*

*mid	Id(mid)	Id(mid)	*fr/rd	Id(rd)	Id(rd)	Id(bk)	Id(bk)
[Seg]	[σ']	[σ']	[Seg]	[σ']	[σ']	[Seg]	[σ']

500	100	100	500	100	100	100	100
-----	-----	-----	-----	-----	-----	-----	-----

2.3.2 The intermediate stages

I ran the GLA through many different learning cycles, to get a sense of the stages that this learner tends to pass through. In each trial, I adopted OTSoft's default re-ranking plasticities, (beginning at 2, and ending at 0.002). The first real re-ranking of markedness and faithfulness occurs after about 700 trials, at which point ranking values are typically as below:

19) Some ranking values after 700 trials

	*mid	Id(mid) [Seg]	Id(mid) [σ']	Id(rd) [Seg]	Id(bk) [Seg]	*fr/rd	Id(rd) [σ']	Id(bk) [σ']
(a)	320.0	287.0	287.0	234.4	235.0	230.6	166.8	168.2
(b)	339.1	260.9	260.9	234.2	236.7	229.1	189.2	203.1
(c)	304.7	295.3	295.3	238	234	228	218	166

The crucial change in this grammar is that *front/rd is now ranked equally with or below general Id-rd and general Id-back. The output distributions table below in 20) shows that this re-ranking has increased the number of options for treating front rounded vowels, and reorganized their frequency:

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

20) The output distributions after 700 trials⁸

target	output	(a)	(b)	(c)
(i) 'kepa	'kipa	100%	100%	99.8%
	'kepa			> 1%
(ii) 'pœki	'puki	5%	> 1%	1.4%
	'pyki	9.1%	4%	
	'pyki	85.9%	95.8%	98.4%
	'pœki			> 1%
(iii) 'lypi	'lupi	> 1%	> 1%	1.4%
	'lipi	9.1%	4%	
	'lypi	85.9%	95.8%	98.6%
(iv) 'pity	'pitu	> 1%	> 1%	1.4%
	'piti	9.1%	4%	
	'pity	85.9%	95.8%	98.6%

In BCD terms, the ranking that this stage roughly matches is:

21) *mid >> Id(rd), Id(bk) >> *front/rd >> Id(rd)-σ', Id(bk)-σ', Id(mid)-both

Unsurprisingly, it is the general Id-round and back constraints that have gotten above Id-round(σ'), because they assign more Ws in (19). And note again that both Id-mid constraints have the same ranking value in every run.

After about 1000 trials, the learner has gotten to a third stage of ranking:

22) Some ranking values after 1000 trials

	Id(mid) [Seg]	Id(mid) [σ']	*mid	Id(rd) [Seg]	Id(bk) [Seg]	*fr/rd	Id(rd) [σ']	Id(bk) [σ']
(a)	302	302	298	236	236	228	176	162
(b)	302	302	298	238	234	228	186	180
(c)	302	302	298	236	234	230	200	192

⁸ At the end of each run of the GLA, OTSoft tested its final state grammar with a 1000 trials on each input. The percentages I give in these output distribution tables reflect the results of those tests.

The crucial change at this stage is that *mid has gotten far enough below Faith that mid vowels are no longer being made high. Along with the continued demotion of *front/round, the learner has now reached a state where very few errors are being made:

23) *The output distributions after 1000 trials*

target	output	(a)	(b)	(c)
(i) 'kepa	'kipa	2.9%	2%	2.7%
	'kepa	97.1%	98%	97.3%
(ii) 'pœki	'puki	> 1%	1.9%	1.4%
	'pyki	> 1%		7.5%
	'pyki	2.9%	2%	2.1%
	'pœki	96.7%	96.1%	89%
(iii) 'lypi	'lupi	> 1%		8%
	'lipi	> 1%	1.9%	0.9%
	'lypi	99.6%	98.1%	91.1
(iv) 'pity	'pitu	> 1%		8%
	'piti	> 1%	1.9%	0.9%
	'pity	99.6%	98.1%	91.1

2.3.3 The end-state grammar

For this learner, the third stage at 1000 trials is pretty much the end-state ranking; none of the crucial rankings that its values embody are going to get revised. To make this perfectly clear:

24) *After 50,000 trials*

	Id(<i>mid</i>) [Seg]	Id(<i>mid</i>) [σ']	* <i>mid</i>	Id(rd) [Seg]	Id(bk) [Seg]	* fr/rd	Id(rd) [σ']	Id(bk) [σ']
(a)	304	304	296	238	238	224	202	144
(b)	304	304	296	238	236	226	216	152
(c)	304.1	304.1	295.9	238	236	225.4	212	196

25) *The output distributions after 1000 trials*

target	output	(a)	(b)	(c)
(i) 'kepa	'kepa	100%	100%	100%
(ii) 'pœki	'pœki	100%	100%	100%
(iii) 'lypi	'lypi	100%	100%	100%
(iv) 'pity	'pity	100%	100%	100%

Since no more errors are being made, these ranking values are therefore final. The comparable BCD final grammar is one that includes these two rankings:

- 26) a) Id(mid)-both >> *mid
 b) Id(rd), Id(bk) >> * front/mid >> Id(rd)-σ', Id(bk)-σ'

2.4 Summarizing the results

2.4.1 The superset grammar: mid vowels

The most important thing we've seen is that *crucial* stringency relations between faithfulness constraints cause the GLA to choose superset grammars. When presented data like 27), the GLA and (my) BCD algorithms get the two different rankings in 28):

27)

winner ~ loser	*mid	Id(<i>mid</i>) [Seg]	Id(<i>mid</i>) [σ']
képa~kípa	L	W	W

- 28)a stabilized GLA ranking: Id(*mid*), Id(*mid*)-σ' >> *mid
 29)b my BCD ranking: Id(*mid*)-σ' >> *mid >> Id(*mid*)

The problem with the GLA-acquired ranking is that it assumes mid vowels are permitted *everywhere*, having only observed them in the stressed syllable context. Thus

the GLA learner will faithfully parse an input with an unstressed mid vowel like [kipe] – precisely the kind of form we wanted our learner to rule out:

29) *The over-generation of the GLA ranking*

/kipe/	Id(mid)	Id(mid)-σ'	*mid
(i) ^{σ'} kipe			*
(ii) kipi	*!		

2.4.2 The ‘Anti-Paninian ‘Ranking: front rounded vowels

A second difference between the rankings that this simulation demonstrated is that errors like 31) below give different rankings from the BCD learner as well:

30)

winner ~ loser	*front/ rd	Id(bk) [Seg]	Id(bk) [σ']	Id(rd) [Seg]	Id(rd) [σ']
lýpi~lipi	L	W	W	W	W
píty~píti	L	W		W	

- 31)a stabilized GLA ranking: Id(rd), Id(bk) >> *rd >> Id(rd)-σ', Id(bk)-σ'
 33)b my BCD ranking: Id(rd)-σ', Id(bk)-σ' >> Id(rd), Id(bk) >> *rd

The difference between these rankings is the position of the positional faithfulness constraints: BCD will install them high when it can, where the GLA doesn't promote them any higher than it has to.

It is not immediately clear which of these end-state grammars should be preferred – because without alternations, it is not clear how we could tell whether adults learning such a language choose one ranking over the other. However, the GLA's quick

promotion of general faithfulness constraints also has interesting but not altogether helpful consequences for its *stages* of acquisition. This is the focus of the next section.

3. Intermediate stages and the Specific-F >> General-F bias in the GLA

A series of studies – Curtin and Zuraw (2001), Boesrma and Levelt (2000); Levelt and van der Vijver (2004) – have proposed using the GLA to generate the intermediate, often overlapping stages seen in children's developing grammars, specifically with reference to the Fikkert/Levelt corpus of Dutch phonological acquisition.

First: Levelt and colleagues connect the order of syllable shape acquisition in Dutch to the lexical frequency of syllable types, and show how the GLA can therefore model these acquisition stages. The constraint set they use – while good for illustrating the GLA and their point – is rather idealized: in particular, their grammar has a monolithic general faithfulness constraint interposed with syllable-shape markedness constraints of various degrees of specificity:

- 32) *Constraints from Levelt and van der Vijver*
Markedness: Onset, NoCoda, *ComplexOnset, *ComplexCoda
Faithfulness: Faith

3.1 The Specific F stages that require the ranking bias

As we saw in the previous section, however, a more articulated set of faithfulness constraints causes problems for the GLA model. The frequency of faithfulness violations always gets a *general* faithfulness constraint promoted either as fast or faster as anything specific to it. As a result, the GLA does not predict intermediate Specific-F stages like the

ones discussed in chapter 3, sections 1.3 and 2.3. Substituting complex onsets for the round vowels in the simulation above, this means that the GLA does not predict the intermediate French stages found by Rose (2000) and Kehoe and Hilaire-Debove (2003) – this is schematized again below:

33) *How the GLA misses the Specific F stage of French complex onsets*

- a) *Initial state:* *ComplexOnset >> Max(Seg)-(σ'), Max(Seg)
Observed intermediate state: Max(Seg)-(σ') >> *ComplexOnset >> Max(Seg)
Target state: Max(Seg) >> *ComplexOnset
- b) *Observed winners:* bá blá bablá blabá
- c) *GLA Constraint movement, created by errors at initial state:*
 *ComplexOnset: demoted by every word with *any* complex onset, i.e.:
blá bablá blabá
- Max-Seg: promoted by every word with *any* complex onset, i.e.:
blá bablá blabá
- Max-Seg(σ'): promoted by every word with *a stressed* complex onset, i.e.:
blá bablá
- d) *Upshot:* *ComplexOnset and Max(Seg) fall and rise at the same rate, while Max(Seg)-σ' rises slower
- e) *The GLA's first new stage:*
 Max(Seg) >> *ComplexOnset >> Max(Seg)-σ': ...saving *all* complex onsets
- f) *The mismatch between (e)'s ranking and the observed intermediate stage:*

	/blablá/	Max(Seg)	*ComplexOnset	Max(Seg)
	(i) babá	**!		**
<i>intermediate stage winner</i>	(ii) bablá	*!	*	*
<i>GLA's winner</i>	(ii) \varnothing blablá		**	

3.2 **The Specific F stages that don't require the bias: Curtin and Zuraw (2001)**

An interesting result, however, is that GLA *can* create intermediate stages in which specific-faithfulness constraints play a role, even though they rank below General-Faith. The Curtin and Zuraw (2001) simulation of Dutch syllable truncation provides such an example.⁹

In their simulation, Curtin and Zuraw provided a GLA learner with schematic lexical items (meaning 2-to-4 syllable strings with one main stress each), at frequencies that approximate the CELEX-based Dutch lexicon. Below I list the Markedness and Faithfulness constraints they used, and their stringency relations:¹⁰

34) *Curtin and Zuraw (2001) constraints*

- | | |
|-------------------------------|-------------------------------|
| (a) <u>Markedness</u> | <u>Faithfulness</u> |
| All-Ft-L | Max-PitchProm (=Max-Stress-σ) |
| All-σ-L | Max-FinalProm (=Max-Final-σ) |
| Parse-Syll | Max-σ |
| FtBin ¹¹ | |
| (b) <u>M-stringency</u> | <u>F-stringency</u> |
| All-Ft-L | Max-Stress-σ and Max-Final-σ |
| <i>is less stringent than</i> | <i>is less stringent than</i> |
| All-σ-L | Max-σ |

In keeping with the M >> F bias (which they point out was crucial to get their early stages to look realistic), they started all M and F constraints with ranking values of 500 and 100 respectively. The general result was that, as in Fikkert's diary study, this simulated learner went through four stages:

⁹ My thanks to Kie Zuraw for discussing this data when I didn't understand what I was saying about it yet.
¹⁰ Readers may remember that I discussed this intermediate stage in chapter 3 section 4.3, without using the constraint All-σ-L.
¹¹ Curtin and Zuraw actually use the constraint 'FootMax', which says "Feet are maximally disyllabic."

- 35) Stage 1: Truncation to one syllable
- Stage 2: Truncation to two syllables (one foot)
- Stage 3: Necessarily one or two feet (no syllables unparsed)
- Stage 4: Unfooted syllables also allowed

Comparing the constraint relationships in 34), and the rankings for these stages, the GLA did precisely what we've seen it does: demoted the general Markedness constraints fastest, and also the promoted the general Faith constraints. Given what I have said above about the GLA's over-zealous promotion of general faithfulness, it is therefore important to note that this learner's stage 2 – truncation to a single foot – *did* preserve syllables in privileged positions (stressed and final syllables).

To see why, I show Curtin and Zuraw's first re-ranking – stage 2 – in 36) below. As expected, the most general markedness constraint, All-σ-L, has gotten below the most general faithfulness constraint, Max-σ:

- 36)a FtBin, Parse-σ, All-Ft-L >> Max-σ >> All-σ-L >> Max-stressed-σ, Max-finalσ

Together, the top three Markedness constraints ensure that every output is no bigger than a disyllabic foot; I collapse all three into the single constraint "OneFoot":

- 36)b "OneFoot" >> Max-σ >> All-σ-L >> Max-stressedσ (and final-σ)

Given that OneFoot is undominated – outputs are going to be no bigger than two syllables, so any input with more than two input syllables must violate general Max-σ. In such a case, the specific Max constraints play the tie-breaking roles even though they are low-ranking:

37) *The GLA-style stage 2: the role of undominated Markedness*

/S ₁ W ₂ W ₃ /	"One Foot"	Max-σ	All-σ-L	Max-stressed-σ	Max-final-σ
(i) (S ₁ W ₂)W ₃	*!		***		
(ii) (S ₁ W ₂)		*	*		*!
(iii) \varnothing (S ₁ W ₃)		*	*		
(iv) (S ₂ W ₃)		*	*	*!	

In this ranking, it doesn't really matter that All-σ-L is between the general and specific faithfulness constraints¹². The respective ranking of specific and general Max is moot in cases where higher-ranked constraints force violation of the general faith constraint.

This analysis of Curtin and Zuraw's stage 2 lets us see why the GLA does not extend to all Specific-F stages. Returning to the example of French complex onsets discussed in the previous section: the problem there is the lack of an undominated Markedness constraint to play the role of 'OneFoot'. The only Markedness at hand requires *no* complex onsets, and so the Specific-F >> M >> General-F ranking is crucial to protect those in stressed syllables.

38) *The intermediate French stage (repeated)*

/blablá /	Max-Seg (Stressed-σ)	*Comp Ons	Max-Seg
(i) blablá		***!	
(ii) blabá		*	*!
(iii) \varnothing bablá		*	*
(iv) babá	*!		*

¹² This is what makes this ranking different from the Anti-Paninian rankings Prince talks about, where it is crucial for other ranking reasons that General-Faith >> M >> Specific-Faith be true. See also chapter 2 section 7.3 for discussion of acquiring true Anti-Paninian rankings.

To get a GLA-style Specific-F stage here, we would need an undominated constraint that wanted *no more than one* complex onset in every word, so that the specific faith constraint could choose its placement:

39) *The GLA version of the intermediate French stage, with a spurious constraint*

/blablá /	<i>OCP-complex onset</i>	Max-Seg	*Comp Ons	Max-Seg (Stressed-σ)
(i) blablá	*!		**	
(ii) blabá		*	*	*!
(iii) bablá		*	*	
(iv) babá		**!		*

So far as I know, such an OCP constraint is not empirically supported. But the larger point here is that deriving Specific F stages in GLA learning relies on the existence of independent markedness pressures to make most of the ranking decisions. This requirement will surely not be met for all such attested stages, so the GLA will not be able to derive them all.

3.3 Interim Summary

Sections 2 and 3 demonstrated that a Specific >> General Faithfulness bias remains necessary in the GLA. The next section considers how that bias should or could be implemented.

4. Persistent biases, contingent biases, and the GLA

The need for the specific >> general faithfulness bias in GLA learning has also been noted recently by Hayes and Londe (2006). In part of their learning simulation of Hungarian vowel harmony, they use a GLA algorithm which includes a ranking bias for

high-ranking IO-Ident[back]-Root >> IO-Ident[back]. They point out in a footnote that general faith climbs too high without such a bias.

In the BCD discussion of chapter 1, ranking biases came in two flavours. First were what I call “definitional” biases, whose effects can be read off the constraints themselves: Markedness >> Faith, OO-Faith >> IO-Faith, and Specific-Faith >> General-Faith (of the language-independent nature.) There are at least two ways in which the GLA could incorporate persistent ranking biases of the definitional sort.

One option would be to assign different plasticities according to both constraint type and direction of re-ranking. To enforce a continued preference for ranking A >> B, we would promote A a lot when it prefers winners, but only demote it a little when it prefers losers, and we do the reverse for B (promote a little but demote a lot).¹³

A more direct way is what Hayes (200X)’s OTSoft manual calls A Priori Rankings. Hayes describes the OTSoft implementation of a priori rankings as follows:

- 40) “OTSoft implements a priori rankings for the Gradual Learning Algorithm as follows: it minimally adjusts the initial ranking values so that any two constraints that are ranked a priori are at least *x* units apart, for some value of *x*. Then, as it incrementally adjusts the ranking values of the constraints, it monitors the a priori rankings so that they continue to be enforced by at least a distance of *x* ranking values or greater. The default setting of *x* is 20, which is very close probabilistically to being an obligatory ranking.” (Hayes, 200X: 21)

Imposed in this way, a priori rankings have the effect of moving constraints *independent of the learning data* in the case that another constraint is getting too close. This approach can thus create a persistent ranking bias for any definitional ranking bias we like.

¹³ Thanks to Joe Pater for discussion of this potential approach.

However, chapter 1 was also concerned with what might be called “calculable” biases – language-dependent Specific-F >> General-F, as well as Prince and Tesar’s (and Hayes’) principles for choosing the right IO-Faith constraint to install in each stratum. These ranking biases were in fact more like ranking *principles*: given a certain set of IO-Faith constraints that prefer a particular set of winners, they determine the safest bet for ranking in the next stratum. It is worth re-stressing that we cannot hardwire all Specific >> General-F relations into the learner because of contingent stringency, and that morphological categories can in principle be implicated in such stringency relations.

The twin problems for including calculable ranking biases in the GLA are (a) how to calculate them and (b) what ranking consequences they should have. The first problem arises because the GLA does not store its errors, so it does not have a Support to draw Context Tables from. Of course, the real-life GLA learner must be learning not just phonotactics but also a real language, and thus a lexicon – so, we could simply say that contingent faithfulness relations are calculated from all the stored URs in the lexicon (with the continued assumption of the Identity Hypothesis whereby inputs are identical to observed winner outputs.)

But there is also the larger question of what these ranking biases actually do to the GLA’s re-rankings. To be concrete, we can first assume our GLA learner is equipped with A Priori rankings for all definitional specific-to-general faith relations – and further that every time our learner goes to promote any W-assigning faithfulness constraints it first consults the lexicon and checks for a contingent specificity relationship between that constraint’s contexts and all others.

And if the GLA learner discovers such a specific-to-general relationship? What should it do? We could first say that it adds that relationship to its list of A Priori rankings – and so if the current error only provides evidence for promoting the less specific constraint, it should nevertheless move up the more specific one a healthy 20 points above the more general one, as instructed in 40) above.

But what if the learner has discovered this specificity relation *after* our constraints have already gotten too high? For example: imagine that our GLA learner is trying to learn Language 1 from chapter 2 §4.3.1.1, where initial syllables are a special case of stressed syllables, and only initial syllables can contain mid vowels. If both of these Ident[mid] constraints have already climbed above *mid in the ranking – it will do no good to notice at some later point that Ident[mid]-σ1 should have been ranked *a priori* above Ident[mid]-σ’. What the learner needs to do to get out of this superset grammar is to demote Ident[mid]-σ’ below the relevant markedness constraint *[mid]. This is a way in which the GLA’s disconnect between errors and ERC rows seems to cause it real problems: the GLA has no notion of demoting below a certain constraint – it only moves constraints along its numerical scale, without reference to the (ordinal) position of any other constraints. And as it stands, it is not clear how the GLA could incorporate the necessary reasoning into its method of constraint re-ranking.

5. A first problem with not storing errors: winner misparses

The rest of this chapter returns to the issues raised in chapter 2 about the need for stored errors (beyond any specificity calculations), and their absence in the GLA.

In some ways, memoryless-learning has its benefits: one way is in dealing with occasional noise in the learning data. Since the GLA's response to each individual learning datum is conservative: if a particular error is in some way wrong, it will have a negligible effect on the end-state ranking. And since the GLA does not remember its errors, there is no sense in which the grammar remains responsible for this error as it continues learning.

However, chapter 2 presented two learning situations in which a memory for learning errors was crucial. One problem, addressed in section 4.1 below, is that unlike occasional noise, persistently misleading learning data in the form of winner misparses will drive the GLA to adopt a superset grammar. Once such grammars have been learned, the GLA's lack of stored errors prevents its recovery even once the misparses are fixed (cf. the Support-based treatment of this problem in chapter 3 §5.2)

Another problem, mentioned briefly in chapter 2 §2.2, is that not storing errors prevents the GLA from seeing fundamental inconsistency between the errors it is making. One place this inconsistency will be present in natural languages is if the target grammar includes exceptionality. What I demonstrate here is that the effect of an exceptional grammar will be the same as a *variable* one – and that revising lexical entries cannot provide the whole solution to this problem. Sections 5.3-5.5 discuss this issue, including the more complex treatment of exceptionality in Zuraw (2000) and Hayes and Londe (2006).

5.1 The GLA's treatment of misparsed winners

To demonstrate the problem, I return to the example of coda devoicing and the mis-syllabification of winners discussed in chapter 2 §4.2. The core of the problem

presented there was the difficulty in learning a correct grammar of voicing neutralization in coda position, when relying on learning data whose voiced onset segments are sometimes mistakenly syllabified as codas:

41) *The target grammar (repeated from chapter 2 ex. 45)*
Ident[voice]-Onset >> *VoicedObs >> Ident[voice]

42a) *Learning onset voicing with the right ERC (chapter 2 ex. 46)*

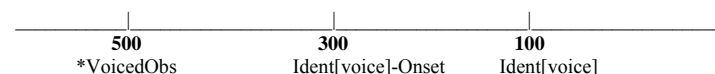
winner ~ loser	*VoicedObs	Ident[voice]-Ons	Ident[voice]
[ká.bla] ~ [ká.pla]	L	W	W

42b) *Learning onset voicing with the misparsed ERC (chapter 2 ex. 47)*

winner ~ loser	*VoicedObs	Ident[voice]-Ons	Ident[voice]
[ká b .la] ~ [ká p .la]	L	e	W

As we saw in section 2, the GLA does not correctly rank specific faithfulness constraints over general ones like Ident[voice]-Onset >> Ident[voice] without a ranking bias. Suppose therefore that we equip the GLA with an initial Specific-F >> General-F ranking bias à la Hayes and Londe (2006) (putting aside the issue of language-specific cases for the sake of argument.) In this case, we will therefore start our learner off with a ranking like 43):

43) A hypothetical GLA initial state, with both M >> F and Spec-F >> Gen-F:



The learner who knows that words like [dábla] are parsed with a medial complex onset will make errors corresponding to the correct learning ERC row in 42a). In GLA

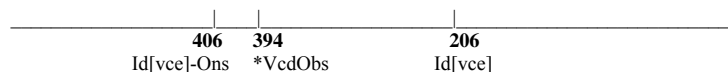
learning, this error will cause a small demotion of *Voiced Obs and a small promotion of both Ident[voice] constraints:

44) The re-ranking effect of the right learning ERC:



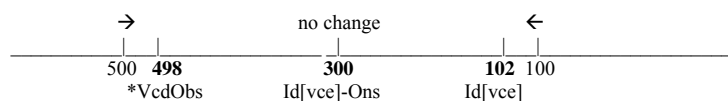
Repeated errors like 44) will eventually promote Ident-Onset *above* *VcdObs; once Ident-Onset has gotten high enough, no more errors will be made, and thus the learner will have found the right end-state grammar:

45) The right final grammar (hypothetical values):



But what about the learner who misparses the relevant winner as [ká**b**.la], and so makes the misleading error in 42b)? The GLA will react to this error with a small demotion of *VcdObs and a small promotion of just the *general* Ident[vce] constraint:

46) The re-ranking effect of the *wrong* learning ERC:



What will this learner's end-state grammar be? For the learner who persists in their syllabification misparse of words like 'kábla', the final grammar will be the wrong one: *VcdObs will continue to sink and general Ident[vce] will continue to rise until their ranking flips and errors stop being made. With these initial ranking values (and the OTSoft default plasticity settings, as in section 2 above), this re-ranking takes about 400 trials; further refinement continues until the grammar reaches a ranking where general Faith is high enough above Markedness that no errors are being made, practically speaking:¹⁴

47)a) Ranking values after 400 trials	b) Ranking values after 50,000 trials
302.660 Ident[vce]	306.000 Ident[vce]
300.000 Ident[vce]-Onset	300.000 Ident[vce]-Onset
297.340 *VcdObstruent	294.000 *VcdObstruent

As we've seen, the GLA does not have a mechanism for unlearning things it has already learned. As a result, the GLA learner's only hope is to learn the right complex onset syllabification of [ká**pl**a] before the crucial re-ranking has happened. If for example only 250 trials have gone by, the re-ranking will not yet have undone the initial rankings:

48) Ranking values after 250 trials

362.236 *VcdObs
300.000 Ident[vce]-Ons
237.764 Ident[vce]

¹⁴ Using an A Priori ranking, rather than an initial ranking bias, will not help this learner at all – it will get Ident[vce]-Onset ranked higher than Ident[vce], but the learner still won't stop making errors until general Ident is above Markedness.

If at this point the learner started making errors using the correct winner parse, Ident[vce]-Onset will start assigning Ws, and *both* Ident constraints will now begin to climb as in 44). We can simulate this second stage of GLA learning by taking the numbers in 48) as initial ranking values, and feeding the GLA the correct input file repeated below:

49) *Learning onset voicing with the right ERC*

winner ~ loser	*VoicedObs	Ident[voice]-Ons	Ident[voice]
[ká. bla] ~ [ká. pla]	L	W	W

In this situation, the learner *will* get to the right end-state grammar. Since both Ident constraints are now rising at the same pace, the more specific Ident constraint will overcome *VoicedObstruent before the general constraint, and this will be enough to stop making errors. In this case: errors stop being made after about 75 trials, and the ranking stabilizes correctly:

50) a) Ranking values after 75 trials	b) Rankings values after 50,000 trials
333.264 Ident[vce]-Ons	336.000 Ident[vce]-Ons
328.972 *VcdObs	326.236 *VcdObs
271.028 Ident[vce]	273.764 Ident[vce]

What this example shows is that the GLA's robustness to persistent misparses is only a function of the *number* of misparsed errors – and, when using initial ranking biases, the speed with which the relevant constraints move in the hierarchy.

5.2 **Winner misparses and markedness: the same problem**

It should be noted that the misparse danger outlined above exists independent of faithfulness and its specificity relations. Suppose that CON included two markedness constraints: *Pharyngeal, and *Pharyngeal-affix, and that the learner was attempting to acquire a language in which pharyngeal consonants only occur in roots but not affixes. This means the target grammar would be as in 51):

51) *The target grammar*
*Phar-Affix >> Ident-Phar >> *Phar

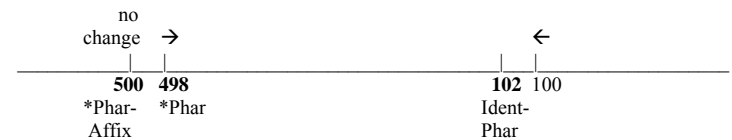
With the right ERCs, learning this ranking presents no problem for the GLA. Assuming again the M >> F initial ranking values, the learner will at first make mistakes on pharyngeals and thereby create ERCs like in 52) below (roots are underlined):

52) *The right ERC for the root-pharyngeal language*

winner ~ loser	*Phar-Affix	*Phar	Ident-Phar
<u>ʔabat</u> ~ <u>gabat</u>	e	L	W

As a result only the general *Phar constraint will be demoted:

53) *The GLA re-ranking resulting from 52)*



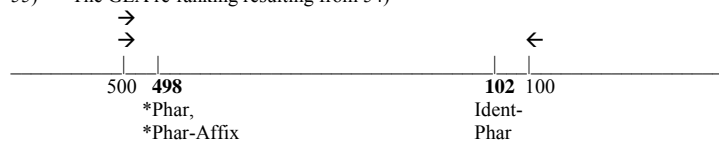
Trouble will arise, however, if the learner happens to misparse words like “ʔabat” as containing the prefix “ʔa-”.¹⁵

54) The morphological misparse ERC

winner ~ loser	*Phar-Affix	*Phar	Ident-Phar
ʔabat ~ gabat	L	L	W

This ERC will make the GLA demote *both* the general and the affix-specific *Phar constraint:

55) The GLA re-ranking resulting from 54)



This is the same situation as we encountered for the syllabification case in the previous section. If this representational misparse lasts long enough, the affix-specific *Phar constraint will eventually get re-ranked below Ident-Phar:

56) *The final-state grammar, learned from incorrect ERCs like 54)*
 Ident-Phar >> *Phar-Affix, *Phar

¹⁵ A couple remarks in defense of the claim that learners might misdiagnose part of a root instead as an affix. The first is that some spontaneous cases of this sort have been cited in the literature – e.g., the exchange: Parent: “Behave!” Child: “I *am* have!” A reasonable counterargument is that this misparse occurred precisely because English *has* a free-standing ‘be’ morpheme, whereas my *Pharyngeal-Affix example requires that the child has mistaken something for an affix when the language crucially does *not* have one. While I do not know of any strictly comparable cases in the literature on morphological acquisition, the most plausible scenario I can think of comes from a language like Hebrew, with both non-concatenative morphology as well as prefixes. The suggestion would be that the non-concatenative properties make children sensitive to vowel patterns, independent of their associated consonants, and that this could transfer over to prefix recognition. Thus, the learner might mistakenly decide on the basis of a series of prefixes and clitic-like things that prefixes were of the format [Ce] (e.g. Hebrew prepositions [be-/le-/ke-/me- as well as verbal pronouns te-/ye-/ne-). So when faced with a root that began [ʔe...] (in Hebrew prime that contained such pharyngeal-initial roots, that is), the learner could choose the relevant morphological misparse.

And even once the learner has learned the correct parse for “ʔabat”, and determined that it does not contain an affix – it is too late for the GLA learner to do anything. Errors are no longer being made, so there is no error-driven incentive to re-rank any of these constraints. And without a memory for its errors, the GLA learner has no way to wonder whether alternative explanations for marked structures like pharyngeals are available.

6. Exceptions and end-state variation

This section looks beyond the stage of pure phonotactic learning, to the later stages where the Identity Hypothesis is abandoned and the learner has adopted unfaithful mappings between inputs and outputs. The goal here is not to explicate how unfaithful inputs are learned, but to consider the behaviour of the BCD and GLA learners once these correct, unfaithful inputs have been chosen.

Among the various grammatical patterns the learner must cope with, I focus on grammars that encode exceptionality and variation. These two patterns are discussed in tandem here because learning theories of both numerical and ordinal OT have sometimes conflated them, though not necessarily to positive effect.

The upshot of this discussion is that the GLA is inherently better suited to handling variation – as is well known – but that the BCD learner is in fact better suited to learning exceptionality. I therefore discuss a way that the BCD could be used to learn variation (extending work by Pater, to appear), and also discuss how the learner of Hayes and Londe (2006), which relies in part on the GLA, would learn exceptions through a non-GLA mechanism. I also emphasize how any learner’s success in discovering the

right patterns of exceptions or variation will derive from their ability to compare stored errors in something like the Support.

6.1 The GLA's treatment of exceptionality

The Gradual Learning Algorithm's approach to variation in phonological development is to make variation an inherent property of the OT grammatical system. Because constraints can overlap in their distribution to varying degrees, the GLA can build a grammar that produces variation between different optimal outputs with probabilities that closely match the attested frequency of variants. The Ilokano example of Boersma and Hayes (2001) shows that the GLA can mirror cases of free variation in this way, by getting just the right constraints ranked in a clump.

A necessary difference between grammars, however, is the split between free variation and grammatical *exceptions*, which show similar patterning among outputs but require different constraint rankings.¹⁶ This difference is not something the GLA is equipped to detect. Without storing errors, the learner can't know whether they have evidence for one pattern of variable grammatical behaviour, or for two different patterns of categorical grammatical behaviour.

What I demonstrate in the rest of section 6.1 are two languages with different input-output mappings, that will nevertheless teach a simple GLA the same variation grammar in both cases. I will then return in §6.2 to more recent GLA on exceptions – the proposals of Zuraw (2000) and Hayes and Londe (2006), which use the GLA in more nuanced ways and do learn exceptionality.

¹⁶ See Pater (to appear), p. 6, for explicit discussion of this difference.

6.1.1 Two languages and their codas

The first example language is the Variable Coda language: one in which every coda in the lexicon is variably deleted in a way controlled by the grammar. To simplify from the real-life grammatical conditions of variable coda deletion attested in natural languages (segmental context, stress, word position, etc.), I will describe the variable coda language as treating *every* coda in citation form exactly the same: retaining it two thirds of the time, and deleting it the other third. This is the grammar that the GLA will learn correctly:

57) *The Variable Coda language*

a) the lexicon (inputs)	b) the outputs
/pak/	[pak], 67% of the time; [pa] 33%
/blag/	[blag] 67% of the time, [bla] 33%
/gri/	[gri] 100%
/tro/	[tro] 100%

In contrast, we might also have an Exceptional Coda language: in this case, most lexical items with codas retain them on the surface in citation form, but an exceptional class of lexical items lose them.¹⁷ (To support the claim that these exceptional words have underlying codas, we can say that they do surface in some non-citation environments – e.g., when preceding a vowel-initial word.)

Unlike the previous case, the exceptional coda language varies between coda faithfulness and coda deletion on an item-by-item basis. As 58) shows, some input codas in a one-word utterance are *always* preserved, while others are *always* deleted:

¹⁷ I call this the exceptional class only because it will be less frequent in the hypothetical lexicon. Depending on the analysis of exceptionality adopted, it could be that the lexical items that *retain* their codas would be the ones given special treatment and thus perhaps the exceptions, even though they are more numerous.

58) *The Exceptional Coda language*

a) the lexicon (inputs)	b) the outputs
/pak/	[pak]
/pa/	[pa]
/blag/	[blag]
/bla/	[bla]
/grip/ - exceptional	[gri]
/tro/	[tro]

This is the language that the GLA will ultimately treat as another instance of free variation (although c.f. Zuraw (2000), Hayes and Londe (2006) – see section §6.2.)

6.1.2 Learning the variable coda grammar

During phonotactic learning, both of these languages will look the same to the learner. Some lexical items will have codas, and some will not, and nothing else will be revealed. This is schematized below:

59)

a) <i>I-O mappings assumed in phonotactic learning</i>	b) <i>Real Inputs variable codas</i>	<i>exceptional codas</i>
/pak/ → [pak]	/pak/	/pak/
/pa/ → [pa]	/pak/	/pa/
/blag/ → [blag]	/blag/	/blag/
/bla/ → [bla]	/blag/	/bla/
/gri/ → [gri]	/gri/	/grip/ - exceptional
/tro/ → [tro]	/tro/	/tro/

To learn any of the bold inputs in 59b) above, the learner has to get past the phonotactic learning stage – a process that I have so far not dealt with in either the GLA or BCD contexts. For present purposes, I will assume that the learner has been augmented so as to learn variation among lexical inputs. This could be done, in principle, by determining that two outputs ([pak] and [pa]) have the same meaning, attempting to

collapse them into one lexical entry, and then deciding that /pak/ is the right input choice because CON includes rankings that map /pak/ to both outputs but no ranking that maps /pa/ to both (i.e. there is no constraint that prefers final k-epenthesis.)¹⁸

I will therefore grant that our learner has adopted the new input-output error pairs in 60) below. Since we now have more than one kind of error, the GLA's behaviour will depend in part on the frequency of their occurrence. In the Variable Coda language, every input with a coda retains it two thirds of the time, and loses it the other third. This is reflected in the frequency column that I have added into the table below; OTSoft mirrors these frequencies in generating errors to feed the GLA:

60) *New ERCs for the variable grammar (as handled by the GLA)*

correct input	winner ~ loser	frequency	NoCoda	Max	Dep
(i) /pak/	pak ~ pa	66.7%	L	W	e
(ii) /blag/	blag ~ bla	66.7%	L	W	e
(iii) /pak/	pa ~ pak	33.3%	W	L	e
(iv) blag/	bla ~ blag	33.3%	W	L	e

To model post-phonotactic learning, I gave this input data in 60) to the GLA. Here is what OTSoft turned out:

61) *Ranking values for the variable coda grammar after 50,000 trials*

new ranking	NoCoda	Max	Dep
(a)	238.2	231.8	230
(b)	238	231.6	230
(c)	240	230	229.5

¹⁸It is not clear to me how the GLA learner would work this out. Since the GLA does not reason about rankings in e.g. the way BCD does, I do not know how it could determine whether CON provides a stable grammar that maps /pa/ to [pak] or vice versa. To my knowledge, this is an unresolved issue.

(d)	239.2	231.6	229.2
-----	-------	-------	-------

62) *Output distributions for the variable coda grammars*

input	output	ranking (a)	ranking (b)	ranking (c)	ranking (d)
/pak/	☞ pak	74.1%	64.6%	63.3%	79.6%
	☞ pa	25.8%	35.4%	36.7%	20.4%
	paka	> 1%			

In short: the GLA has learned this variable coda grammar correctly. One could get the frequencies closer to the targets by with some fine-tuning – e.g. changing the schedule of constraint plasticities – but for our purposes it is enough to note that this ranking allows codas more often than not, and that any lexical coda is up for deletion between a fifth and a third of the time.

6.1.3 Learning the exceptional coda grammar

As in the previous section, the learner of the Exceptional Coda language must first do some morpho-phonological processing and change its lexical inputs to realize that something other than the identity map between inputs and outputs is necessary. Here, the learner must find a way from alternations to notice that some lexical items with codas nevertheless surface in citation form without them, getting to a new set of ERC rows like those below:

63) *New ERCs for the exceptional grammar*

correct input	winner ~ loser	NoCoda	Max	Dep
(i) /pak/	pak ~ pa	L	W	e
(ii) /blag/	blag ~ bla	L	W	e
(iii) /grip/	gri ~ grip	W	L	e

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

This input set was constructed so that the ratio of regular to exceptional coda forms is 2:1 – that is, two thirds of the lexicon’s words with codas retain them, while the other third lose them. But recall that this language differs from the variable coda one in that no single lexical item varies: /pak/ always retains its coda in this context; /grip/ always loses its coda. And this lexical regularity is what the GLA cannot track – all it can know is that two thirds of its errors look like 63i), and that the other third look like 63iii). So when I again gave the GLA these three new errors in 63), the results looked just like those in the previous section:

64) *Ranking values for the exceptional coda grammar after 50,000 trials*

initial rankings	new ranking	Dep	Max	NoCoda
62a)	(a)	238.9	231.2	230
	(b)	238.7	231.2	230.1
62b)	(c)	238.4	231.8	229.8
	(d)	238	231.9	230

These rankings produce variable codas in *all* lexical items. The table in 65) below shows the frequency of coda retention and deletion: it is of course the same for each lexical item, regardless of which form always wins in the target language (indicated by the ☞)

65) *Output distributions for the exceptional coda rankings*

input	output	ranking (a)	ranking (b)	ranking (c)	ranking (d)
(i) pak	☞ pak	68.1%	65%	74.3%	71.7%
	pa	31.8%	35%	25.7%	28.2%
(ii) blag	blag	68.1%	65%	74.3%	71.7%
	☞ bla	31.8%	35%	25.7%	28.2%

6.2 Learning exceptions: GLA-related approaches

The previous section showed that a straightforward implementation of the GLA will treat both the variable and exceptional coda grammars the same – in both cases it will learn a variation grammar. I have illustrated this point primarily to add to the arguments against the GLA’s lack of stored errors – because without stored errors that categorically do or don’t follow a generalization, a language with exceptions looks just like a variable one.

There is at least one proposal in the GLA-related literature that is built especially to handle grammars with exceptions. This is the model of Zuraw (2000), which I will refer to here as the Full Listing model for reasons that will become clear.

In the Full Listing model, the GLA is used to learn the stochastic ranking of constraints that govern lexical subregularities.¹⁹ After learning with the GLA, such constraints overlap and so produce variable results: but in Zuraw’s system these constraints are ranked low enough that they can only have any effect on the choice of output for *words that do not have a lexical entry* – that is, new coinages, nonce words or unfamiliar words heard uttered by another speaker. In contrast, the shape of every input in the lexicon (whether deemed regular or exceptional by the linguist) is protected by high-ranking I-O lexical faithfulness constraints.²⁰

In fact, this model exploits the GLA’s ability to learn a variation grammar from categorically regular and exceptional inputs in a very clever way. What the GLA does for the Full Listing learner is tinker with the ranking of the constraints that prefer and

¹⁹ C.f. the ordinal OT approach of Pater and Coetzee (2005), designed for a related kind of learning off the lexicon.

²⁰ As well as a constraint UseListed, which ensures that learners use their stored lexical forms as inputs, rather than creating new inputs for known words and letting low-ranked constraints affect their output.

disprefer particular processes, in tune with their lexical frequency, while steadily moving faithfulness up the hierarchy. Once IO-faith gets high enough, errors stop being made and learning is done. The resulting grammar is one in which, for everyday production purposes, IO-faith makes all the decisions. The learner has encoded all forms in the lexicon, both regulars and exceptions, and they will always be faithfully produced. But at the bottom of the hierarchy resides the variation grammar that the GLA learned. This ranking is the locus of speaker’s knowledge about statistical regularities in their lexicon, and it also perpetuates lexical tendencies across generations of learners by driving the frequency with which *novel* words enter the language as regular or exceptional.²¹

What this Full Listing model gives up, however, is the goal that the grammar be responsible for the phonological regularity or exceptionality of lexical items. The violation profile of each winner was used in learning to build the low-ranking stochastic grammar – but in the final adult grammar that profile is not important because IO-faith keeps forms identical to their stored forms. As a result, the Full Listing learner’s grammar is not built to be restrictive, and does not rule out a Rich Base.

Hayes and Londe (2006)’s analysis of Hungarian stem-controlled vowel harmony adopts some key assumptions from the Full Listing model, but these authors also aim to build a restrictive grammar independent of the lexicon. They do so by proposing a two-step learner, splitting the learner’s task into two consecutive parts: first, to acquire a restrictive phonotactic grammar of what is possible and impossible in Hungarian, and second, to acquire the statistical regularities of where the possible patterns of harmony occur, and how often. To accomplish the first task, they use a non-GLA learning algorithm: either BCD or Hayes’ LFCD (Hayes, 2004; see my discussion in chapter 2

²¹ I have done rather short shrift to this approach; see Zuraw (2000) chapter 2 for the whole story.

section 2), which shares with BCD the crucial properties of stored errors and the acquisition of ordinal rankings.

To accomplish their second learning task, the two-step learner uses the GLA, but with two large caveats. First, they use the constraint rankings acquired in phonotactic learning *as a priori rankings for GLA learning*; second, they remove faithfulness *entirely* from the constraint set that the GLA works with. Thus, all the GLA is allowed to learn is the relative frequency of various harmonic and disharmonic vowel sequences, via the frequency of their markedness violations.

The two-step Learner of Hayes and Londe also adopts the Full Listing assumption – namely that all derived lexical items are stored fully composed with their suffixes. In their analysis, each Hungarian stem is stored with the allomorph(s) that it is observed to take: the front suffix, the back suffix, or both in the case of stems that the authors call ‘vacillators’. In this way the two-step learner’s grammar is indeed responsible for deriving forms in the usual generative way. On the one hand, the ranking learned by BCD or LFCD in their first acquisition step ensures that stems stored with only one suffix will optimally keep that suffix in the output. On the other hand, vacillators are provided to the grammar as inputs with both suffix allomorphs, and since high-ranking IO-faith cannot choose between allomorphs²² the low-ranking stochastic rankings learned by the GLA will choose between them (just as they do in Zuraw’s approach to nonce words.)

Although Hayes and Londe do not discuss this brand of exceptionality: with the assumption of full allomorph listing and a generous notion of allomorphy, their model seems able to capture the exceptional coda grammar from section 6.1.3 above. However,

²² See Kager (1999), Wolf (2005) on this assumption that inputs with two allomorphs map fully-faithfully to outputs with either allomorph.

it does so precisely because its set-up prevents the GLA from playing a role. In the first stage of phonotactic learning, the learner will determine that codas are possible and so acquire the ordinal ranking Faith >> NoCoda (just as in §6.1.3.) Upon discovering that words like [gri] come in fact from inputs like /grip/ (presumably from alternations), the learner must know to store *both* of these outputs as allomorphs of this root – thus, the learner’s inputs will be as in (66) below:

- 66) *The exceptional coda language in a Hayes and Londe-like system*
- | | |
|-----------------|-----------------------------------|
| a) the inputs | b) the outputs (in citation form) |
| /pak/ | [pak] |
| /pa/ | [pa] |
| /blag/ | [blag] |
| /bla/ | [bla] |
| / {grip, gri} / | [gri] |
| /tro/ | [tro] |

With this lexicon, regular inputs with codas will retain them (67a) and exceptional inputs with two allomorphs will lose them via low-ranking NoCoda (67b):

- 67) a) *Deriving regular coda preservation...*
- | | | |
|-----------|-----|--------|
| /pak/ | Max | NoCoda |
| (i) ☞ pak | | * |
| (ii) pa | *! | |
- b) *... and exceptional coda deletion*
- | | | |
|-----------------|-----|--------|
| / {grip, gri} / | Max | NoCoda |
| (i) grip | | *! |
| (ii) ☞ gri | | |

One outstanding question is the extent to which this approach’s brand of allomorphy can handle all the attested types of lexical exceptions (for a sample of real cases see Pater, to appear.) What is clear from this example, though, is that the Two-Step learner of Hayes and Londe acquires this exceptional grammar but uses a re-ranking algorithm by using a BCD-like ranking, and not the GLA.

6.3 Learning variation without the GLA: a BCD approach

Wee have just seen that in contrast to the GLA, a Biased Constraint Demotion learner equipped with a Support can indeed learn an exceptionality grammar. As discussed in chapter 2 §2.2 Pater (to appear) uses Inconsistency Detection and Biased Constraint Demotion to successfully learn a grammar that encodes exceptionality using lexically-indexed constraints (see also Kager, in press; Flack, to appear.) In a related vein, Pater and Coetzee (to appear) also use a similar BCD approach to learn a grammar that encodes lexical subregularities; see also Becker, 2006.

On the other hand, the BCD learner has no extant mechanism for acquiring a variation grammar. Recall that the variable ESL learner I proposed in chapter 2 is variable only as long as it is learning – once all the necessary errors have been added to the final Support and the Error Cache is cleared for the final time, there are no way to cause variation because there is only one ranking. More generally, it is not necessarily clear what an end-state variation grammar should look like in an ordinal OT system, regardless of how it should be learned (on the end-state question, see Nagy and Reynolds, 1995; Inkelas, Orgun and Zoll, 1997; Anttila, 1997, 2002; Pater and Werle, 2003; Pater to appear.) Since the GLA is inherently good at learning such grammars, this chapter's general critique of the GLA model demands some discussion of how an ordinal OT grammar might be made to capture end-state variation – and also learn it.

The approach to variation in Error-Selective Learning in chapter 3§5 started from the claim that variation in an ordinal OT system must be derived by switching between multiple *grammars* – unlike a GLA-style grammar, which can generate different *rankings* to use in each run of EVAL from a single grammar. To re-use the example of the coda

variation grammar in §6.1.2 but now in ordinal OT terms: the speaker of this language will have to sometimes use a ranking in which NoCoda >> Max, and other times use a ranking in which Max >> NoCoda.

Along these lines, one well-known approach to variation in ordinal OT is to use partially-unranked constraints, as proposed by Anttila (1997, 2002); the version I will adopt here is the (2002) model. In this approach, grammars consist of sets of specified constraint rankings which are consistent with all of the language's variation, and with all remaining constraint orderings left unspecified. This means that the speaker of the variable coda language (using just these three constraints) has a grammar in which the only specified ranking is NoCoda >> Dep, and the ranking of Max is left unspecified. Every time the speaker uses the grammar, they will randomly pick a full ordering of the constraints, which means that Max will get ranked either above or below NoCoda:

68) *The variable coda grammar in the partially-unordered constraints model (adapted from Anttila 2002)*

- | | | | |
|----|---------------------|-----------------------|---|
| a) | The grammar: | specified rankings: | NoCoda >> Dep |
| | | unranked constraints: | {Max} |
| b) | The full orderings: | | NoCoda >> Dep >> Max
NoCoda >> Max >> Dep
Max >> NoCoda >> Dep |

(continued on the next page)

c) The variation these fully-ordered rankings create:

(i) first two rankings:

(ii) third ranking:

/pak/	NoCoda	Max	Dep	/pak/	Max	NoCoda	Dep
(i) pak	*!			(i) pak		*	
(ii) pa		*		(ii) pa	*!		

There are some known difficulties with this model,²³ but does provide us with an explicit model of ordinal OT variation with which we can ask the relevant learning question: How could a BCD learner come to adopt the grammar in 68a)?

Pater (2005) provides a proposal for learning variation with the T/S algorithm which, like with everything else in this dissertation, proceeds by examining the properties of its Support. The suggestion in Pater (2005) for how the learner could conclude the need for the partially-unordered rankings of a variation grammar is to notice that the same input has created two or more winners – that is, *if it detects inconsistency among errors with identical inputs*. Such a Support is built below, adapted from the errors fed to the GLA back in table 60):

²³ As pointed out by Pater (to appear), one issue with using partially-unranked constraints to model exceptionality is that at one level it equates variation and exceptionality, just as the GLA does. What we have already seen that this makes the wrong predictions about a truly exceptional grammar – and Pater (to appear)’s more general argument is that exceptional and variable phonological patterns can easily exist one without the other and so require a grammar (and a learner) that knows the difference between them. In any event, I have presented only the bare bones of both the theory and the dissent here – see the Antilla and Pater references given above for a fuller view of the debate. See also Inkelas, Orgun and Zoll (1997) for a related but somewhat different view of partially-unranked constraints to handle variation, though not exceptions. For another issue, see footnote 25.

69) *Errors for the BCD learners of the variable coda grammar*

input	winner ~ loser	NoCoda	Max	Dep
(i) /pak/	pak ~ pa	L	W	e
(ii) /pak/	pa ~ pak	W	L	e
(iii) /blag/	blag ~ bla	L	W	e
(iv) blag/	bla ~ blag	W	L	e

(Note two important things about this Support: first, that the frequency of each variant and its associated error is being ignored in this discussion,²⁴ and second that this learner is at a state where it has already determined the correct underlying representations, including the unfaithful ones as in rows (ii) and (iv), just as was assumed in the previous GLA discussion.)

On the basis of a Support like (69), Pater (2005) proposes that the learner acquire two fully-specified rankings, each of which covers one of the two inconsistent ERC rows. Below, I spell out a slightly different way of learning from this inconsistency, relying on the partially-unranked constraints we have already used to capture variation in this section.

The BCD learner will fail to find a ranking for the Support in 69) as soon as it gets started: it cannot install a single constraint, because its Markedness constraint NoCoda prefers losers, and neither faithfulness constraint prefers only winners. To check whether its inability to find a grammar is due to variation in the data (instead of exceptionality, incorrect representational assumptions or anything else), the learner might now consider each set of errors in the Support *that have the same input*, and attempt to

²⁴ In the Antilla (2002) model used in 68), the frequency with which each variant is chosen is a function of how many rankings of the unordered constraints choose that variant. The tableaux in 68c) show that, given this constraint set and grammar, coda deletion should be chosen 66% of the time, coda faithfulness should be chosen 33% of the time, and no other frequency of variation is predicted. While more constraints could be introduced to prefer one variant more often than the other, the connection between variants and their frequency is clearly dealt with much more elegantly in the GLA system, where just two constraints can create many different patterns of proportional variation as a function of the distance between their ranking values.

find a ranking for just that set. In the Support in 69) above, this means grouping together the first two ERC rows, and the last two ERC rows – and right away we can see that both subsets will again be found inconsistent, because they each contain conflicting information about the ranking of NoCoda and Max:

70) *The Support built e.g. only for the input /pak/ is still inconsistent*

input	winner ~ loser	NoCoda	Max	Dep
(i) /pak/	pak ~ pa	L	W	e
(ii) /pak/	pa ~ pak	W	L	e

In the face of this inconsistency among errors for the same input, the learner can at least be sure that if these two input-winner pairs are right, then grammar it is learning contains variation. In the view I have adopted here, the effect of this discovery must be that the learner chooses a set of constraints to be unordered in the grammar.

Looking at 70), it is clear that the conflict between NoCoda and Max are responsible for the inconsistency. How will the learner see this? Recalling the Antilla (2002) model of variation, the method I propose is that the learner first builds rankings for each individual ERC row in the single-input Support (like 70), and then builds a grammar which specifies all and only the constraint orderings common to all these ERC-specific rankings.

If we take this first step for the two errors in 70), we will nearly get the two ERC-specific rankings below:

71) *Building rankings for each ERC row variant*

a) the first ERC from 70):

Input	winner ~ loser	NoCoda	Max	Dep
(i) /pak/	pak ~ pa	L	W	e

which builds this grammar:

Max >> NoCoda >> Dep

b) the second ERC from 70):

Input	winner ~ loser	NoCoda	Max	Dep
(ii) /pak/	pa ~ pak	W	L	e

which builds this grammar:

NoCoda >> Max, Dep

To take the second step, the learner can compare each constraint pair and only specify those whose ordering is the same in all the ERC-specific ranking. As spelled out in 72) below, this will build a grammar like in 68), with just NoCoda >> Dep specified:

72) *Building a variation grammar from the rankings in 71)*

- a) NoCoda and Dep: NoCoda >> Dep in 71a)
NoCoda >> Dep in 71b) ... specified in the grammar
- b) NoCoda and Max: Max >> NoCoda in 71a)
NoCoda >> Max in 71b) ... so left unspecified
- c) Max and Dep: Max >> Dep in 71a)
Max, Dep in 71b)²⁵ ... so left unspecified
- d) Final grammar: NoCoda >> Dep
{Max}

Admittedly: this procedure for collapsing ERC-specific rankings into a Antilla (2002) variation grammar works fine for a system with only 3 constraints, but would no doubt require refinement when scaling up to learning grammars with more constraints

²⁵ In this comparison, I have chosen to treat the constraints left unranked within a stratum by the BCD algorithm as different from those crucially ranked. But note that the BCD algorithm puts Max and Dep in the same stratum here *not* because it is crucial that they be unranked – recall Step 4 of the algorithm in chapter 2, which simply dumps all remaining IO-faithfulness constraints at the bottom of the hierarchy in one stratum.) If we instead chose to adopt the specified ranking in this situation – here, Max >> Dep – we would make the same predictions for the kinds of variation the language shows (coda deletion vs. faithfulness), but different predictions about the frequency of each variant (see previous footnote.) I leave this interesting difference for further work.

(and possibly more complicated interactions.) Nevertheless we have at least seen that an explicit procedure can be defined for the BCD learner to build a partially-ordered variation grammar.

Assuming that this procedure or something like it will allow the learner to get from a Support like 69) to the grammar in 72d), there must now also be some revision to the normal workings of the learner. In all subsequent attempts to feed observed forms through the current grammar, the learner now has a *number* of multiple rankings that all instantiate that current grammar. It is no longer clear how the learner should proceed in parsing observed outputs, and determining when an error has been generated.

Pater (2005) suggests that error generation in a variation grammar could proceed by assuming that if an observed output can be produced faithfully by at least *one* of the rankings consistent with the current grammar, it does not count as an error. However, this technique will be insufficient because it will prevent the learner from seeing any lexical exceptions once a variation grammar has been adopted. Just because /pak/ and /blag/ retain their codas in a variable fashion does not mean that every coda in the language behaves this way; the partially-ordered grammar will always have *one* ranking that does each, but it will never notice on its own that e.g. /blak/ always retains its coda while e.g. /pag/ never does (see again Pater to appear.)

At this point, the best way to learn using a variation grammar and the Error-Selective BCD learner must be left as an open issue. But given what has come before, I suggest that a likely approach to noticing the difference between exception and variable inputs will again come from the BCD learner's analysis of the Support and its associated rankings, as well as the range of ERC rows that share common inputs.

Overall, this section's discussion has demonstrated that an enriched analysis of the Support can at least provide the BCD learner with the knowledge that phonological variation is part of its target ordinal OT grammar. The method by which this learner moves from an inconsistent set of ERCs with a common input to the right variation grammar, and how adopting a variation grammar at an intermediate stage affects the further course of BCD learning, remain open questions.

6.4 Summary

This section has compared the GLA and BCD approaches to the discovery of exceptionality and variation in target grammars. With respect to these two areas, both algorithms have their strong and weak points. On the one hand, BCD learning is better suited to diagnose and encode categorical exceptions, because unlike the GLA it notices that the conflicting ranking information it is receiving comes from different inputs. On the other hand, GLA learning (of numerical OT grammars) is better suited to capturing variation because unlike an ordinal OT learner it can easily respond to conflicting evidence by gradually overlapping the ranking of relevant constraints.

7. Chapter Summary

This chapter has discussed the Gradual Learning Algorithm, an alternative approach to the OT learning questions treated in chapters one and two. While the GLA is inherently a learner that goes through intermediate stages, I have argued that it is not the best way to model phonotactic learning. I have demonstrated its restrictiveness problems with specific faithfulness relations, arguing that they cannot be fully treated using ranking

biases, and also pointed to its inability to produce some Specific-F intermediate stages. I have also argued that the choice not to retain its errors, in contrast to BCD's storage of errors in the Support, makes the GLA unable to recover from early winner misparses and the superset rankings they cause.

Finally, I have discussed the extent to which the GLA and the BCD learners can each handle grammars with exceptional forms or variation within forms. To return to one of the broader themes of this dissertation: this discussion has served to highlight that whether our eventual grammar uses ordinal or numerical rankings, it is still crucial that the learner keep a memory of its errors to determine the true nature of the pattern it is attempting to learn.

CHAPTER V

TESTING FOR THE HIGH-RANKING OO-FAITH BIAS

1. Introduction to the chapter

This chapter returns to the ranking bias proposed in Hayes (2004) and McCarthy (1998): the preference for high-ranking paradigm uniformity constraints, formalized here using Output-Output Faithfulness (OO-Faith; Benua, 2000). I will expand on the arguments discussed in chapter 2 that this bias provides a mechanism for preventing the acquisition of superset grammars, and also suggest following Hayes that it finds independent support in some children's innovative application of paradigm uniformity in the literature (§1.3).

One interesting aspect of this approach are its predictions (exemplified in §2.2 - §2.3) about the behavior of the phonological grammar at the point of *morphological* acquisition: that is, the point at which derived words are first decomposed into multiple morphemes and phonological patterns become attributable in principle to the demands of OO-Faith.

In sections 3 through 5 of this chapter, I report on an experiment which tested these predictions, using an artificial language 'wug test' (Berko, 1958) to test 4-year-old children's production of marked consonant clusters in two different morphological environments. The results (§5) support the claim that children prefer to repair clusters in ways that satisfy OO-Faith at the expense of other Markedness and Faithfulness pressures, and I provide some analyses (§6) of individual participant's productions and the associated rankings that match theoretical predictions.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

In section 7, I step back from the experimental details to consider the theoretical implications of these findings, including their positive contribution to the artificial learning literature, and some independent support for the OO-faith bias. In section 8, I discuss a possible experimental confound and its entailed questions for future research.

2. The OO-faith ranking bias and phonotactic learning

2.1 The role of OO-faith in enforcing restrictiveness

Chapter 2 (§3.2.3) made the point from McCarthy (1998) that a high-ranking OO-faith bias will make sure that paradigms are kept uniform and non-alternating unless evidence points to the contrary. Hayes (1999/2004) makes the same restrictiveness point for OO-faithful allophony, in languages where the normal distribution of some allophone is overridden just to keep a morphological paradigm uniform. The famous example he discusses is the interaction of flapping and Canadian raising (CR) in some dialects of English, already introduced in chapter 2 §3.2. To recall the facts: CR is purely allophonic in monomorphemic words: raised [ʌɪ] appears before voiceless obstruents, as in 'write' [ɹʌɪt], while [aɪ] appears elsewhere as in 'ride' [ɹaɪd]. However, derived forms with a base vowel [ʌɪ] exceptionally retain their raised quality even before a voiced flap, as in 'writer' [ɹʌɪrəɹ], *[ɹaɪrəɹ]. In the OO-faith analysis summarized in chapter 2, 'writer' contains a raised diphthong because it is faithful to the vowel quality of its morphological base 'write', whose raised vowel is allophonically-conditioned.

Hayes' learning argument is this: at the point that a learner encounters the word 'writer', how does he or she account for the presence of marked [ʌɪ] before a voiced flap? The child who does not yet know that 'writer' is derived from 'write' cannot explain

‘writer’s raised vowel with OO-faithfulness – instead, the error below (using the constraints from chapter 1) simply suggests that the [ʌir] sequence is licit:

1) *The morphologically-naïve error caused by ‘writer’*

/ʌɪrəʌ/	Ident[lo]-OO	*ait	*ʌi	Ident[lo]-IO
(a) [☞] ʌɪrəʌ				*!
(b) ʌɪrəʌ			*	

This error tells the BCD learner that IO faithfulness to any properties of raised vowels (here, Ident-[lo]) must rank above the general markedness constraint that disprefers raised vowels in the pre-flap context (see 2ii below).

2) *Initial Support during phonotactic learning, with no morphological relations*

input	winner ~ loser	Ident[lo]-OO	*ait	*ʌi	Ident[lo]-IO
(i) /ʌɪt/	[ʌɪt] ~ [ɹɪt]	e	W	L	W
(ii) /ʌɪrəʌ/	[ʌɪrəʌ] ~ [ɹɪrəʌ]	e	e	L	W

Since OO-faith to vowel quality does not prefer any losers, BCD can rank it at the top of its ranking – however, it also does not prefer any winners, so it doesn’t resolve either of our errors. And while contextual markedness *does* explain the raised vowel in roots like ‘write’ in (2i), the ERC row in (2ii) can only be explained by IO-faithfulness, and thus BCD comes up with the ranking in 3):

3) *The resulting ranking that BCD learns: a superset grammar*
 Ident[lo]-OO >> *ait >> **Ident[lo]-IO** >>> *ʌi

As the bold emphasizes, this ranking is a superset grammar because it preserves input raised diphthongs in *all* contexts.

Hayes points out using data about the empirical timeline of acquisition that children do not know enough morphology early enough to avoid the superset trap illustrated in 2) and 3). And he therefore suggests that an OO-Faith bias offers part of a solution to this learning trap, if it is construed as a persistent bias that can return the grammar to a more restrictive state once morphological learning has occurred. When the learner realizes that ‘writer’ includes the base ‘write’, OO-faith constraints now prefer the raised vowel in the derived form. Imported into the present framework, this realization will mean that the Support’s ERC row resulting from 1) will be updated to look like 4):

4) *Revised Support, post-morphological learning (bases underlined)*

input	winner ~ loser	Ident[lo]-OO	*ait	*ʌi	Ident[lo]-IO
(i) /ʌɪt/	[ʌɪt] ~ [ɹɪt]	e	W	L	W
(ii) /ʌɪt/ + əʌ/	[ʌɪrəʌ] ~ [ɹɪrəʌ]	W	e	L	W

Giving this revised Support to my BCD learner gets us the right ranking. In its first stratum, the Main Routine ranks OO-faith and resolves the error in (4ii); in the next stratum it ranks *ait and so resolves (4i). With all errors resolved, the algorithm ranks the rest of its constraints according to the M >> F bias, and the right ranking has been found:

5) *The ranking that BCD learns from 4): the correct grammar*
 Ident[lo]-OO >> *ait >> *ʌi >> **Ident[lo]-IO**

I will return to the acquisition timeline of phonotactics vs. morphology, and the consequences for the nature of our ranking biases, in section 7.

Now that we have reviewed the end-state argument for a BCD-style ranking bias for OO-faith: what does this bias predict about stages of acquisition in the present system? The rest of this chapter considers this question in some detail.

2.2. Predictions for stages of acquisition

It was shown in (1) through (3) above that during pre-morphological learning, the OO-faith ranking bias causes BCD to continually re-install OO-faith constraints at the top of the hierarchy, even though they have no effect in resolving errors or driving mappings. However – once some morphological bases have been learned, these high-ranking OO-faith constraints will suddenly kick in and begin to assign violations. At this point, they should therefore begin to drive a new kind of error: enforcing paradigm uniformity on paradigms that are *not* OO-faithful in the target.

The following sections illustrate this, using an example that foreshadows the experimental test of the prediction to come.

2.2.1 The target: an OO-unfaithful language

Consider a language where a particular Markedness constraint is freely violated; to use an example relevant to the experiment to come, imagine this constraint is Agree(Voice), (Lombardi, 1996, 1999), which requires obstruent clusters to agree for voicing. In our toy grammar, the faithfulness constraints that conflict with Agree[voice] are:

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- | | | |
|----|----------------|---|
| 6) | Id[vce]-IO | Output segments must match their input correspondents for voicing |
| | Id[vce]-OO | Output segments <i>in a derived form</i> must match their <i>base</i> correspondents for voicing ¹ |
| | Id[vce]-Ons-IO | Output segments <i>which are syllabified as onsets</i> must match their input correspondents for voicing |

Before processing any errors, our BCD learner ranks these constraints as in 7):

- 7) Id[vce]-OO >> Agree[voice] >> Id[vce]-Ons-IO >> Id[vce]-IO

To see the effects of OO-Faith at different stages, we will consider the optimal outputs for two lexical items with different morphology:

- 8) *The toy lexicon*
 (a) /zɪtʃdɪn/ simple word
 (b) /wʌtʃ + dəl/ derived word – morphological base /wʌtʃ/

2.2.2 OO-faith kicks in at the initial state

How would the initial stage grammar treat these two words is this base ‘wutch’ had already been identified? Ignoring the implausibility of this order of acquisition, let us describe the grammars effect on our two-word lexicon in (8). With the ranking in (7), voicing mismatches would be repaired, but repaired differently depending on whether the cluster spans a base/affix boundary (in /wʌtʃ + dəl/) or not (in /zɪtʃdɪn/). In simple forms, the repair would remain the same as above: voicing is protected in onset, so it is changed in coda:

9) *Initial state, once post-morphology: simple word*

/zitʃdm/	Id-[vce]-OO	Agree[vce]	Id-Ons[vce]-IO	Id-[vce]-IO
zitʃdm		*!		
zitʃtm			*!	*
☞ zɪdʒdm				*

In derived forms, however, coda voicing would now be protected by OO-Faith as part of a morphological base, so voicing would change in *onset*:

10) *Initial state, post-morphology: correctly analyzed /wʌtʃdəl/*

/wʌtʃ + dəl/	Id-[vce]-OO	Agree[vce]	Id[vce]-Ons-IO	Id-[vce]-IO
wʌtʃdəl		*!		
☞ wʌtʃtəl			*	*
wʌdʒdəl	*!			*

2.2.3 OO-faith kicks in at an intermediate stage

Returning to Hayes' point about learning timelines – it is probable by the time OO-faith begins to assign violations, much phonotactic learning will already have occurred. However this does not necessarily mean that the learner has reached the target ranking.

Imagine, for example, that the base 'wutch' were to be identified an intermediate stage between initial and final. Since the constraint set we are considering here includes just one Markedness constraint, the only possible intermediate stage is the Specific-F ranking in 11) below. This learner has promoted just a specific faithfulness constraint,

¹ This is a rather simplified definition of OO-Faith – see footnote 12 of chapter 1 – but the simplification does not affect the analysis here in any substantive way.

Ident(voice)-Onset, above Agree(voice), on the basis of some error that ESL has added to the Support (see §6.1 for what that error would look like.)

11) Id[vce]-OO >> Id[vce]-Ons-IO >> **Agree[voice]** >>> Id[vce]-IO

Note, though, that this re-ranking has not been sufficient to tolerate voicing mismatches in general, because Agree(voice) still ranks above general Ident(voice).

What if morphology was learned at this point – what outputs would now be optimal for our two-word lexicon? Like the previous stage, the ranking in 11) chooses different outputs for this cluster, dependent on the morphology. In simple forms, the cluster is still repaired by a change in coda voicing:

12) *Intermediate stage, simple word*

/zitʃdm/	Id-[vce]-OO	Id-Ons[vce]-IO	Agree[vce]	Id-[vce]-IO
zitʃdm			*!	
zitʃtm		*!		*
☞ zɪdʒdm				*

For the form /wʌtʃ + dəl/, however, OO-Faith blocks voicing change in coda, and onset-specific IO-Faith blocks a repair in onset. As a result, the optimal candidate is the faithful one, in which mid-ranking Markedness is violated and the voicing mismatch survives:

13) *Intermediate stage, derived word*

/wʌtʃ + dəl/	Id-[vce]-OO	Id-Ons[vce]-IO	Agree[vce]	Id-[vce]-IO
☞ wʌtʃdəl			*	
wʌtʃtəl		*!		*
wʌdʒdəl	*!			*

3. The experimental methodology: artificial language learning

3.1. The difficulties in testing for OO-faith in L1 acquisition

The previous chapter laid out the learnability case for high-ranking OO-faith and made predictions about stages of innovative OO-faithfulness in development. However, testing these predictions in natural L1 learning poses some problems.

The largest problem is catching children at the right stage – and being sure that it *is* the right stage. Unlike purely phonological analyses of children’s production, the predicted effects of OO-faith at any stage are also tied to the learner’s representational assumptions about morphology – its bases, relations, paradigms, and the like. Thus, the claim that a particular pattern results from the OO-faith influence is also a claim that children have learned enough morpho-semantics to calculate the right OO-faith relations and enforce their violation profiles in the ERCs added to the Support. In this respect, phonologically-transcribed data from sources like CHILDES will be insufficient in most cases to be sure of a child’s state of morphological awareness.

Furthermore, English morphology does not provide many good testing grounds for such investigation. To see the innovative effects of OO-faith in a developing grammar, a child must have learned a morphological base that undergoes some phonological process or change in its target derived form, so that the child can block that process in their own derived word productions. One of the few good cases in English is flapping, which in the target causes an alternation between base-final [t]s and [d]s and their flapped correspondents in derived forms with vowel-initial suffixes:

14) One potential English case of innovative OO-faith: base-final flapping

a) The non-OO-faithful pattern in the target (bases underlined):

bases	target derived forms	predicted OO-faithful forms
‘wait’ [wa <u>it</u>]	‘waiter’ [wa <u>ɪ</u> tɹ]	‘waiter’ [wa <u>ɪ</u> tɹ]
‘sit’ [s <u>it</u>]	‘sitting’ [s <u>ɪ</u> tɪŋ]	‘sitting’ [s <u>ɪ</u> tɪŋ]
‘need’ [n <u>id</u>]	‘needed’ [n <u>ɪ</u> rəd]	‘needed’ [n <u>ɪ</u> dəd]

In fact, Bernhardt and Stemberger (1998) do report the case of a child who went through this stage (see also the cases discussed in section 7.3.1) However, the biggest potential English cases come the shift of stress in fairly complicated derived words (cf. *cycle* ~ *cyclity*) that the average four year old has probably not heard very often (and whose morphological decomposition they have therefore probably not learned.)

For these two reasons, this chapter investigates the predictions of OO-faith in phonotactic learning using a somewhat novel experimental methodology: artificial language learning. In the next section I introduce this kind of testing method, and then present the child-directed version of the approach that I combined with a standard wug-test to tackle the OO-faith question.

3.2. The artificial language learning paradigm

Artificial language learning is an experimental technique where subjects are trained and tested on novel language material – words, sounds or the like – with the explicit understanding that they are from a language unknown to them. Training is implicit and quick, usually involving brief exposure to the novel language stimuli with little instruction other than to remember as much of the materials as possible. Testing can take a variety of forms, but its overall goal is to investigate what subjects have internalized about the novel language data – or, put differently, what effects the novel

data have had on the subject's pre-existing linguistic knowledge. Usually, the materials are constructed specifically for the purposes of the experiment; although they are frequently constructed to resemble patterns attested in natural language learning, it is not necessary that they be the actual words or inventories of any *one* attested language. In addition, the paradigm allow us to compare the acquisition of *unattested* language patterns to attested ones.

The rationale behind artificial language learning is that it can be used ask questions or test predictions that are hard to test in natural settings, and also provides a way to control for as many confounds as possible. In the recent literature, research in this paradigm has compare adults' acquisition of minimally-different languages to test theories of how particular factors affect the ease or speed of phonological learning: properties such as phonetic naturalness, natural language attestedness, statistical patterns and probabilities, as well as the connection between static phonotactic and productive alternations. Research in this vein includes Esper (1925); Saffran, Aslin and Newport (1996); Saffran, Newport and Aslin (1996); Pater and Tessier (2003), (2005); Pycha et al (2003); Wilson (2003), (2006); Carpenter (2005), (2006); Peperkamp and Dupoux (2006); Morrison (2005). After discussing my own experimental results and their interpretation, I will return in section 8.1 to broader questions about this methodology and its applications.

3.3 The present application

The first crucial difference in this artificial language learning study is that it was with children: English-speaking four year olds who, (presumably) unlike adult subjects,

were still in the process of learning their L 1. To my knowledge, no previous work has used an artificial language learning paradigm of this sort with children. (It should be noted, however, that some infant speech perception research has asked related questions, in testing infants' abilities to acquire novel phonetic categories – e.g. Maye, 2000 – as well as novel lexical representations: Saffran, Aslin and Newport, 1996; Stager and Werker, 1997; Werker et al, 1998; Chambers et al, 2003.)

In the present context, the experimental goal was to test predictions about the effect of morphological discovery on phonological production. To induce morphological learning, I taught them a novel bound morpheme – a plural suffix – through direct comparison of singular and plural forms. To test the effects of newly-learned morphology on production I wanted to induce phonotactic errors, so the materials included marked coda-onset clusters, which were either very low-frequency or absent in English. This design was an attempt to simulate a morphologically-informed but phonologically-novel state, where morphological relations were known (and so OO-faith could be active), but high-ranking markedness constraints were still high enough to induce phonotactic repairs. By virtue of using novel language stimuli, we can be sure that participants were encountering the language's bases and plural suffix for the first time.

One question about this methodology is what grammar children are using when producing novel forms in this experiment; given that the participants were four years old, they had clearly learned most of the basic English phonological system. In most regards, the materials of this experiment were designed to allow for some agnosticism about the extent of English-specific grammar that participants brought to this task – the materials

provided children with marked structures they were unfamiliar with, and so would probably have not yet acquired even outside the experimental context.²

4. Experimental Design

4.1 Experimental predictions

The words of the artificial language taught in this experiment were designed to compare faithfulness to bisyllabic forms in the two morphological conditions exemplified in section 2.2: within a morphologically-simple word, and across the base-affix boundary of a complex form. In my experiment, the Count Plural condition contained words like [wʌtʃ.del], where the first syllable was the singular base and the second syllable was the novel plural suffix; the Mass Noun condition contained words like [zɪtʃ.dɪn], where both syllables were neither base nor affix. As we saw above, grammars that have not completely demoted markedness constraints against a particular marked cluster should repair it differently in these two morphological contexts. This is summarized below:

15) *Two predicted asymmetric patterns of faithfulness, using OO-faith:*

- | | <i>simple word</i> | <i>derived word</i> |
|----------------------------------|-------------------------|--|
| a) initial ranking in (7): | /zɪtʃ.dɪn/ → [zɪdʒ.dɪn] | /wʌtʃ _[base] .dəl/ → [wʌtʃ.təl] |
| b) intermediate ranking in (11): | /zɪtʃ.dɪn/ → [zɪdʒ.dɪn] | /wʌtʃ _[base] .dəl/ → [wʌtʃ.dəl] |

Thus, the experiment was aimed at answering two specific questions about children's cluster productions in the two morphological conditions:

² One important related issue, however, is whether four-year old children have learned anything about English that drives them to be OO-faithful to clusters in the plural context that my experiment relied on. This is an issue I must look into further, and which will affect the design of subsequent experiments.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

- 16) *Prediction 1, with respect to initial syllable codas*
Coda segments should be produced more faithfully in the first syllable of plural nouns than in the first syllable of mass nouns – e.g., more faithfulness to /tʃ/ in /wʌtʃ/ + /dəl/ than in /zɪtʃ.dɪn/

This asymmetry should manifest itself when children have made the morphological connection between base and affix and incorporated it into their phonology. Under both the initial and intermediate rankings, base codas should be protected – either at the expense of onset segments at the initial state, or Markedness violations at the intermediate state – whereas mass noun first-syllable codas should not.

- 17) *Prediction 2, with respect to unfaithful medial clusters:*
Among those tokens whose medial clusters are produced unfaithfully in some way, more of the unfaithfulness should be seen in onset position in the count nouns than the mass nouns.

This prediction is a specific test for the two-repair asymmetry of initial state. In those tokens where high-ranking Markedness has driven an unfaithful repair, the ranking of OO >> IO faith predicts onset repairs for count nouns, where OO-Faith protects the coda, but not for mass nouns where OO-Faith has no effect.

4.2 Materials

Table 23 below shows the representative properties of all the novel words children learned (details of how they learned them in the next section). Count noun singulars were all mono-syllables, of the shape CVC(C); each singular was suffixed with [dəl] to form a bisyllabic plural with the shape CVC(C).[dəl]. Mass nouns were all bisyllabic, of the shape CVC(C).dVC. All bisyllabic forms were initially stressed; all vowels in the second

syllable were lax (of the set ɪ, ɛ, ə) and pronounced as unstressed but not completely reduced. Every effort was made by the experimenter to produce the clusters and their segments similarly in all tokens and contexts: in particular, coda stops were somewhat released (as they might be in very careful English speech). These measures were taken to attempt to make the two morphological types of bisyllabic words forms as prosodically-similar as possible, but to maximize the perceptability between the second syllables of plurals (always [dəl]) and mass nouns (always of the form [dVC].) Every cluster occurred both within a mass noun and across the count noun-plural suffix boundary.

The full set of clusters and items used:

18)

Coda Segment(s)	Cluster	Count	Mass
Stop(s)	b.d	pob (+ dəl)	gɪbdɪt
	g.d	wʌg (+ dəl)	mɒgdəm
	kt.d	lʌkt (+ dəl)	pæktɪm
Fricative-Stop	ft.d	fʌft (+ dəl)	lʌftɪdek
Affricate	tʃ.d	wʌtʃ (+ dəl)	zɪtʃɪm
	dʒ.d	bɪdʒ (+ dəl)	fɒdʒɪt
	bʒ.d	gɪbʒ (+ dəl)	mæbʒɪt
Nasal-Fricative	nf.d	nanf (+ dəl)	gʌnfɪp
	mf.d	namf (+ dəl)	gʌmfɪp

4.3 Methodology

The experiment was presented to the children as a novel language-learning game, in which their task was to learn some “alien” words, spoken by an alien puppet named Bozdim (operated by the experimenter.) Children were taught the alien words by association with pictures of familiar objects: count nouns in singular and plural contexts,

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

and mass nouns in two different containers. The children therefore learned a series of new nouns, as well as a novel plural suffix [dəl].

In training, children heard the puppet produce the alien words each paired with an object, and were encouraged to both imitate and spontaneously reproduce them in conversation. In the second part, the children played a picture-matching game with the puppet, in which the puppet named a picture and asked children to name a matching picture. In this game, children were tested for their pronunciation of the same clusters in bases (the two-syllable mass nouns with medial clusters) vs. derived forms (the plural, suffixed count nouns with clusters created at the morphological boundary.)

4.3.1 Participants

Twelve 4-year old children in the Amherst and Northampton, Massachusetts areas participated in the study. This age group was chosen because four-year-olds typically have fairly adult-like phonological grammars, but have not completely mastered difficult segments and clusters. With respect to morphology, four year olds are also reported to be at the stage of over-generalizing regular morphological patterns – e.g. *foots*, *mouses*; *ranned*, *bringed* – and thus presumably in the throes of productive morphological acquisition.³

4.3.2 Training

Initially, the experimenter (and puppet) presented children with picture one at a time. First, children were asked for the English name for the object, and engaged in short

³ In fact, three of the children whose data is reported here produced English morphological errors of this sort during spontaneous conversation, including “He bringed the chair” and “I runned to the table”.

discussion of the object and its properties – its colour, size, prototypicality, etc. – to get the child focused on the object. Then, the puppet was asked to give the name of the object in his language.⁴ Once the puppet had given the object’s name, the child was encouraged to repeat the name (“Can you say what Bozdim just said?”), and to use it in similar discussion as with the English name before (“Is this a blue wug? Or is it a yellow wug?” or “Does it look like this cup of zitchdin tastes good? Do you think the zitchdin is hot or cold?”)

In this phase, children first learned three words of one noun class – count or mass – and then three of the other. Within a count noun block, participants first learned three singular nouns, and then their corresponding three plurals. Within a mass noun block, participants learned three mass nouns, and then heard the same three again in a different container – a glass of juice, and later a bottle of juice. Half of the participants saw count nouns before mass nouns, and the other half saw mass before count.

The set of materials that children learned in this initial block are given below:

⁴ Ideally, the child would ask the puppet, but in the face of shyness the experimenter would do so instead.

19) *Materials*

<i>Noun class</i>	<i>Morphology</i>	<i>Prosodic shape</i>	<i>Sample words</i>	<i>Sample matching pictures</i>
Count	Three singulars (base)	CVC(C)	[pɒb]	one armchair
			[wʌtʃ]	one pick-up truck
			[nænf]	one flower
	Three plurals (base + suffix)	CVC(C).dəl	[pɒbdəl]	many armchairs
			[wʌtʃdəl]	a fleet of pick-up trucks
			[nænfədəl]	a garden of various flowers
Mass	Three mass nouns	CVC(C).dVC	[gɪbdet]	a glass of juice
			[zɪtʃdm]	a cup of hot chocolate
			[gʌnfdep]	a mug of milk
	Same three mass nouns	(same 3 as above 3)	[gɪbdet]	a bottle of juice
			[zɪtʃdm]	many cups of hot chocolate
			[gʌnfdep]	a carton of milk

4.3.3 **Testing**

Once children had learned three of each words, the experimenter asked the children to play a matching game with the puppet. All twelve pictures seen so far were laid out in front of the child; to play the game, the puppet pointed to one picture and named it for the child. The child would then find the matching picture, and name it for the puppet. In this game, the puppet pointed to one of each of the mass nouns, for the child to match by naming the other, and to each of the *singular* count nouns, for the child to match by naming the *plural*. Thus, testing asked the child to provide six words, all with difficult coda-onset clusters: three underived mass nouns, and three derived plural nouns from singular bases.

After the first game, children were presented another training block as in 19), with six more words, and another testing game was played. (For some children, only two more

words of each category were taught, so that the second round of testing included only 4 words.)

5. Experimental Results

5.1 The data reported

To make any claims about the morphologically-sensitive phonological patterns in the data, we must be able to claim that participants had in fact learned the artificial language's morphology – that is, learned its plural suffix “del”. In order to prove sufficient mastery of /dəl/, I required that participants provide at least one spontaneous token of more than one plural noun, associated with the right plural picture. This criterion eliminated 2 participants, leaving 10 children.

Of the 9 clusters tested, only 5 are included in the final results. Two criteria excluded the others: first, the cluster had to be pronounced unfaithfully in more than 2 tokens; second, it had to have been produced by more than 1 child in both the mass and plural contexts. The first criterion eliminated two clusters (gd, kd) and the second another three (ftd, ktd, b3d), which leaves the following clusters:

- 20) *obstruent/d* *nasal-fricative/d*
 b.d mf.d
 tʃ.d nf.d
 dʒ.d

All results reported are for these 5 clusters and 10 children. All tokens were reported, from what was referred to above as both training and testing – this is simply to get enough tokens.

In addition: plural tokens were only included when the participant produced a second syllable of type dV(C). In other words, the results do not include tokens with English plural affixes (“wutʃez, pobdelz”) or zero morphology (“wutʃ, pob”).

5.2 Testing the predictions

The majority of children's pronunciations were of two types: either faithful, or with reduction in the coda of the first syllable. To first give an impression of the data, table 21) summarizes the general results (variances across the 10 subjects given in parentheses):

21) *Results, across subject and by condition*

	<i>total tokens</i>		<i>faithful codas out of total tokens</i>		<i>unfaithful medial clusters out of total</i>		<i>faithful codas out of total unfaithful clusters</i>	
	#	#	%	#	%	#	%	
plural nouns	87	69/87	0.793 (0.035)	25/87	0.287	9/25	0.36 (0.1711)	
mass nouns	112	56/112	0.50 (0.06)	52/112	0.464	1/52	0.019 (0.006)	
totals	199	125		77		10		

5.2.1 Testing prediction 1

The data in table 22) below allows us to test prediction 1 – that codas in initial syllables should be more faithful in count nouns than mass ones. The table shows the raw number of tokens of faithful coda productions that each subject produced in the two morphological conditions, and also the proportion of all tokens that were coda-faithful in each condition:

22) Proportion of faithful $\sigma 1$ codas by subject and condition

Subject	Mass Nouns			Plural Nouns		
	faithful codas	total codas	% coda-faith	faithful codas	total codas	% coda-faith
C	17	20	0.85	11	11	1
E	4	9	0.444	9	9	1
A2	2	7	0.286	6	6	1
I	4	15	0.267	5	8	0.625
N2	8	12	0.667	9	12	0.75
A3	1	10	0.1	7	9	0.778
A1	3	9	0.333	10	15	0.667
D1	13	17	0.765	3	4	0.75
D2	1	3	0.333	7	10	0.7
N1	3	10	0.30	3	3	1
totals	56	112		70	87	
means			0.5			0.805
variance			0.06			0.024

Summing across all 10 subjects, a one-tailed t-test showed that codas were produced faithfully significantly less often in mass noun clusters, namely in 50% of tokens, than in the plural count noun clusters, where they were faithful 79.3% of the time ($p < 0.01$). Further, a pair-wise t-test, comparing the proportion means for each subject, also shows a significant difference between the lower proportions of faithful first-syllable codas produced in mass nouns compared to the higher proportions in plural nouns ($p < 0.01$).⁵

Thus, prediction 1 seems nicely borne out. This result provides some evidence for intermediate stage rankings, where plural nouns are faithful to both members of medial clusters, but mass nouns are still unfaithful in coda position.

⁵ Statistics were calculated both for all 10 subjects, but also using just the first 8, since the low number of total items for the last two subjects, D2 and N1, might have skewed the proportions. Either way, however, the result is significant at $p < 0.01$.

5.2.2 Testing prediction 2

Table 23) below tests prediction 2 – that derived forms will show not only more faithfulness in their cluster codas, but also less faithfulness in cluster onsets. This result would correspond to an initial state ranking, where clusters are repaired in the *codas* of mass nouns, but the *onsets* of plural ones. To test for such a possibility, I consider again the proportion of faithful codas, but only among those that were unfaithful somewhere in the medial cluster:

23) Proportion of faithful codas in unfaithful medial clusters by subject and condition

Subject	Mass Nouns				Plural Nouns			
	faithful codas	faithful onsets	totals	% coda-faith	faithful codas	faithful onsets	totals	% coda-faith
C	1	3	4	0.25	0	0	0	0
E	0	5	5	0	1	0	1	1
A2	0	3	3	0	0	0	0	0
I	0	8	8	0	0	3	3	0
N2	0	4	4	0	1	3	4	0.25
A3	0	9	9	0	3	2	5	0.6
A1	0	6	6	0	1	5	6	0.1667
D1	0	4	4	0	0	1	1	0
D2	0	2	2	0	0	3	3	0
N1	0	7	7	0	0	0	0	0
totals	1	51	52		6	17	23	
means				0.0192				0.261
variance				0.0063				0.1711

The above table shows that unfaithful clusters were overwhelmingly unfaithful in coda position: only 1/52 mass nouns and 9/25 plural nouns with unfaithful clusters had faithful codas. This is not too surprising, given the privileged faithful status of onsets over codas (as encoded in the positional faithfulness Ident-Onset constraints used through this dissertation.)

Despite this clear tendency to apply repairs in coda, table 23) does provide some support for prediction 2. Across all 10 subjects, a one-tailed t-test assuming unequal variances shows that the mean proportion of coda faithfulness in these clusters is lower for mass than for plural nouns ($t = -2.077$; $p = 0.0322$.)⁶

5.3 Summary of results

The results of the previous section were positive: the data from our 10 children and 5 clusters does show the influence of morphological structure on phonological faithfulness, in the ways predicted by high-ranking OO-faith and the two developmental stages discussed.

While the results in the previous section were described in terms of faithful vs. unfaithful segments, it is clear that children differ in their patterns of repair, and for different clusters. Using table 23) as a guide, it seems that A3 and N1 are both at a stage where the effects of both the initial state and intermediate stages can be seen. The next section considers these two subjects' treatments of particular clusters in the two conditions in-depth, which as we will see provide ranking arguments that match the schematic stages already seen.

6. Rankings in the results

In this section, I focus on a few examples in the data collected with illustrate the particular rankings predicted in section 2.2.

⁶ Across just the first 8 subjects, the difference between the means remains significant with a one-tailed t-test ($t = -2.2131$; $p = 0.0289$).

6.1 A3's cluster voicing: an intermediate ranking

Two nice examples of the predicted stages come from A3's treatment of the cluster /tʃ.d/. First, consider her pronunciation of two words below, with respect to cluster *voicing*.

24) /zɪtʃdɪn/ [zɪdʒdɪn]

25) /wʌtʃ + dəl/ [wʌtʃdʒəl]
c.f. /wʌtʃ/ [wʌtʃ]

The pronunciation [zɪdʒ.dɪn] in 24) suggests that A3 does not tolerate affricate-stop clusters that disagree for voice, and that she can repair this disagreement by voicing the offending coda. The ranking in 26) derives this pattern: Agree(voice) >> Ident(voice) requires that voicing be repaired somewhere, and coda changes in voicing rather than the onset due to Ident(voice)-Onset.

26)

/zɪtʃdɪn/	Agree(voice)	Ident(voice)-Onset	Ident(voice)
zɪtʃ.dɪn	*!		
zɪtʃ.tɪn		*!	*
→ zɪdʒ.dɪn			*

Comparing this tableau to the token in 24) however, we can see that A3's grammar chooses different winners when this cluster is created at a base-suffix boundary: here, voicing surfaces faithfully, violating Agree(voice). The morphological explanation offered here is that the winners preserve the *base's* input voicing in this derived form.

Note that in the plural noun /wʌtʃ + dəl/, the coda affricate is part of the word's morphological base /wʌtʃ/ – and that A3 produces /wʌtʃ/ faithfully, with a voiceless coda. As 27) again shows, ranking the OO version of Id-voice above Agree(voice) predicts this pattern. Id(voice)-OO prevents coda voicing and positional faithfulness prevents onset devoicing in, so the (voicing) faithful candidate wins:

27)

/wʌtʃ + dəl/	Id(vce)-OO	Id(vce)-Ons-IO	Agree(vce)	Id(vce)-IO
(a) wʌtʃtəl		*!		*
(b) wʌdʒdəl	*!			*
→ (c) wʌtʃdʒəl			*	

This result is positive in two respects. First, A3's asymmetrical treatment of cluster voicing mismatches, despite the language's equal tolerance of them, is an intermediate stage ranking of the sort predicted by this gradual learner.

Second, the nature of this asymmetry in A3's grammar lends support to the hypothesized OO-faith bias as well. Without such a pressure, it is not clear why A3 should protect coda voicing – at the expense of the markedness of voicing disagreement – only in forms whose codas form part of a morphological base.

6.2 A3's treatment of coda affricates

Now we will consider two of A3's other tokens, with respect to the affricate [tʃ]:

⁷ A3 also produced the form [wʌtʃʒəl]: while different in its treatment of cluster continuancy, this form also shows the progressive voicing of the winner in 31c) and so is compatible with 31)'s ranking.

28) /zɪtʃdɪm/ (a) [zɪʔdʒɪm]

29) /wʌtʃ + dəl/ (a) [wʌtʃdʒəl]
c.f. /wʌtʃ/ (b) [wʌtʃ]

In both morphological contexts, here A3 repairs some marked aspect of the cluster /tʃ.d/, although she produces coda /tʃ/ on its own faithfully in 29)b).⁸

For present purposes, I assume that the repair in the mass case, [zɪʔdʒɪm], represents the effect of a simple ban on coda affricates. The mass form is repaired by moving the continuant over to onset position, violating low-ranking IO faithfulness to continuancy.⁹

30)

/zɪtʃdɪm/	*CodaAffricate	Ident[+cont]-IO
(a) zɪtʃdɪm	*!	
(b) → zɪʔdʒɪm		**

In the count noun case, the singular base also violates *CodaAffricate, and yet it is faithfully produced. This can be explained as another specific faithfulness effect: i.e. to word-final position.

31)

/wʌtʃ/	MaxSeg-WdFinal	*CodaAffricate	Ident[+cont]-IO
(a) → wʌtʃ		*	
(b) wʌʔ	*!		*

⁸ I will leave aside the spreading of continuancy onto the plural affix in 29)a) although this effect seems ripe for analysis using a Markedness pressure for clusters to agree in continuancy, plus OO-faithfulness to the base.

⁹ I assume that simply deleting the continuant feature is not an option, due to other high-ranking faithfulness constraints.

The crucial test case is A3's treatment of the count noun /wʌtʃ + dəl/: while in this form the coda affricate is no longer at the word edge, it is still produced faithfully. The explanation here is that its deletion is blocked by OO-faith, i.e. Ident[+cont]-OO.

32)

/wʌtʃ _{base} + dəl/	Id[+cont]-OO	*CodaAffricate	Id[+cont]-IO
(a) wʌʔdʒəl	*!		*
(b) → wʌtʃdʒəl		*	*

This ranking in 32) simply demonstrates one preference of OO-faith over IO-faith in A3's grammar:

33) OO-Faith >> Markedness >> (General) Faith
 Id[+cont]-OO >> *Coda Affricate >> Id[+cont]-IO

Again, the choice of two different repairs seems inexplicable without a morphologically-sensitive pressure like OO-faith.

6.3 N's treatment of [mf.d] clusters: an initial state ranking

A similar example comes from N's unfaithful realizations of /mfd/ clusters. N does sometimes produce this cluster faithfully, but looking just at the unfaithful ones reveals another two-repair asymmetry:

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
 Ph.D. dissertation, UMass Amherst

	Input	Output
34)	a) /gumfdin/	[gumfdin]
35)	a) /næmf + dəl/	[næmfəl]
	b) c.f. /næmf/	[næmf]

(Note that I am abstracting away from [p], which N inserts variably for all three of these forms.)

Whatever the markedness problem with a coda [mf] sequence – described here with the purely ad-hoc constraint *Coda[mf] -- N's grammar offers two different ways of solving it. In the mass noun, he retains all the input segments, but changes the place of articulation of the coda nasal from labial to coronal:

36)

/gumfdin/	*Coda[mf]	Max(Seg)-IO	Max(Labial)-IO
(a) gumfdin	*!		
(b) gumfin		*!	*
(d) → gunfdin			**

However – this repair is not applied in 35b) in the plural /næmf + dəl/ *[nænfədəl]. Instead, N deletes a following *onset* segment. On the current story, this is understood as another effect of high-ranking OO-Faith: the labial coda [m] in /næmf + dəl/ is protected by its membership to the base by Max(Labial)-OO, so a different IO constraint must be violated to satisfy markedness:

37)

/næmf+ dəl/	Max(Lab)-OO	*Coda[mf]	Max(Seg)-IO	Max(Lab)-IO
(a) næmfədəl		*!		
(b) → næmfəl			*	
(d) næmfədəl	*!			*

6.4 Summary of analyses

This section has highlighted three cases where the experiment results match the theoretical model. In all three cases, the child produced the same coda-onset cluster differently according to its morphological contexts, remaining preferentially faithful to base material in a derived context just as the OO-faithfulness account predicts.

7. Theoretical discussion

7.1 The intermediate stage, and Error-Selective Learning

The predictions that lead to my experiment were about the kinds of errors OO-faith could induce in a phonotactic learner – both at the initial stage in which markedness is still all-powerful (the ranking in 7), and at an intermediate stage where specific faithfulness alone has overcome markedness (the ranking in 11). This first stage, where Markedness is always obeyed, was a starting assumption of the learning theory I've adopted throughout this dissertation. But we have not yet seen why the Error-Selective BCD learner should go through this particular intermediate stage, whose ranking from 11) I repeat below:

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

38) (repeated from 11)
 Id[vce]-OO >> Id[vce]-Ons-IO >> **Agree[voice]** >>> Id[vce]-IO

To build such a ranking, the Error-Selective learner will have to have added an error to the Support where positional Ident-Ons[voice] prefers the winner so that BCD can install it high. With just the two errors and four constraints of section 2.2, however, this wouldn't happen – in that Error Cache, the only faith constraint doing any work is the general IO-Ident[voice]:

39) *Initial Support during phonotactic learning, with no morphological relations (repeated from 2)*

input	winner ~ loser	Ident[lo]-OO	*ait	* _{AI}	Ident[lo]-IO
(ii)/ɹɹɹɹɹɹ/	[ɹɹɹɹɹɹ] ~ [ɹɹɹɹɹɹ]	e	e	L	W

If, however, we had a slightly more complex CON that included more markedness constraints, it could well be the case that some errors would optimally satisfy Agree-Voice by changing onset rather than coda voice. Such is the case in the hypothetical Error Cache below:

40) *An Error Cache*

winner ~ loser	Id[vce]-OO	*Vcd Velar (= *[g])	*Affricate	Agree [vce]	Id-Ons[vce]-IO	Id-[vce]-IO
(i) zɹtɹdm ~ zɹɹdm	e	e	L	L	e	W
(ii) zɹtgm ~ zɹtkm	e	L	e	L	W	W

If Agree[voice] were now to trigger Error-Selective learning on this Cache: the Markedness criterion of the ESA wouldn't choose between the two errors, because they each violate Agree[voice] and one other Markedness constraint. Thus the Faithfulness

The child Marina had been correctly producing Greek velar and palatal fricative allophones – up until 4;7, when for a few weeks she began defaulting to the velar fricative just where the palatal disrupts the paradigm uniformity of verbal stems:

45) *Marina at 4:7:fricative palatalization blocked by OO-Faith*

	2pl.	'to have'	Target [eçete]	Child [exete]
	2pl.	'to leave'	[fevʝete]	[fevçete]

46) *Marina's grammar*

OO-Max[velar] >> *[xi, ʝi] >> *[ç, j] >> IO-Max[velar]

This data fits neatly with the present theory. The BCD algorithm predicts that Marina should indeed have been installing OO-Max[velar] at the top of her ranking all along, for no reason other than her biases. Below that, the distributional evidence from her phonotactic learning ranked her markedness constraints in the right allophonic order. With the assumption that she has now realized that words like [eçete] have a verbal base /ex-/ , these new errors have emerged under new pressure from undominated OO-faith:

47)a) *Simple words: fricative place decided by *[xi]*

/exete/ (hypothetical)	OO-Max [velar]	*[xi, ʝi]	*[ç, j]	IO-Max [velar]
(i) exete		*!		
(ii) → eçete			*	*

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*.
Ph.D. dissertation, UMass Amherst

47)b) *Derived words: palatalization can be blocked by OO-Max[velar]*

/ex-/base + /ete/	OO-Max [velar]	*[xi, ʝi]	*[ç, j]	IO-Max [velar]
(i) → exete		*		
(ii) eçete	*!		*	*

Hayes also cites Bernhardt and Stemberger's (1998) case mentioned in §3.1, of an English-learning child who, from 2;0 – 3;8, consistently flapped in simplex words (e.g. 'water' [warər]), but only produced base-faithful voiced and voiceless stops in derived words even where the adult phonology required a flap (e.g. 'sitting' [sɪtɪŋ] from base [sɪt] and 'needed' [nɪrəd] from base [nɪd])¹⁰.

Another case of OO-faith innovation comes from Smith (1973)'s seminal diary study of English acquisition by Amahl. Both Macken (1980) and Jesney (2005) point out a paradigm uniformity quirk in Amahl's puzzle-puddle-pickle chain shift. This shift in Amahl's grammar between 2;2 and 2;11 caused coronal stops to become velar before laterals, in both simple and derived words:

48) *Amahl's velarization before laterals: /t,d/ → [k,g]*

'puddle' [pʌgəl]	'gentle' [dɛŋkəl]	'padding' [pægəlɪn]
'turtle' [tʰəkəl] ¹¹	'gently' [dɛŋkli:]	'pedaling' [pɛgəlɪn]

This pattern is a more extreme version of the English ban on tl and dl onset clusters (cf. the syllabifications of ma.tress vs. at.las) – in Amahl's grammar, these sequences are ruled out regardless of syllabic position. For present purposes I simply

¹⁰ See Bernhardt and Stemberger (1998) for their alternative analysis of this data.

¹¹ Amahl was learning a British English dialect that lacks this post-vocalic [ɹ].

adopt the constraint *tl, intended as an OCP constraint that disallows sequences of coronal stops and laterals, and rank this constraint above faithfulness to consonantal place (see Jesney, 2005 for a full treatment of this chain shift in Amahl's grammar.)

49) *Amahl's pattern of pre-lateral velarization*

/pʌdəl/	*tl	Max [place]
(a) pʌdəl		*
☞ (b) pʌgəl	*!	

There is one class of words in which pre-lateral coronals do not become velarized: derived words whose *base* had only the stop and not the following lateral. As shown in 21) below, words like 'tight' surface with their normal coronal stop, and this stop is retained in derived words like 'tightly' even though provide the phonological context for velarization:

50) Velarization blocked *when the base has no velarized segment:*

'hard'	[hɑ:d]	'hardly'	[hɑ:dli]
'soft'	[sɒft]	'softly'	[sɒftli:]
'tight'	[taɪt]	'tightly'	[taɪtli:]

The explanation adopted here is that coronal stop in 'tightly' is required to be OO-faithful to the stop in its base, 'tight'.¹²

¹² An alternative account of this data is that Amahl's pre-coronal velarization was the result of misperception (that he was hearing tɪ as [kɪ] in words like 'gentle' or 'gently'), and that those words whose bases had no l allowed him to perceive the coronal correctly – see Macken (1980).

51) *Adding OO-faith to Amahl's ranking*

/tait/	OO-Max [place]	*tl	IO-Max [place]	/tait + ly/	OO-Max [place]	*tl	IO-Max [place]
☞ (a) tait				☞ (a) taitly		*	
(b) taik			*!	(b) taikly	*!		*

7.3 The persistent OO-faith bias and the GLA

7.3.1 The empirical need for persistent OO-faith

Section 2.1 alluded to the argument in Hayes (2004) that the OO-faith bias must be persistent, given the facts of order of acquisition. The facts that Hayes refers to come from the growing body of experimental work about *receptive* learning in very young children. On the one hand, this work has demonstrated that children have internalized native phonotactic distributions in some sense roughly by the age of 8-10 months (e.g. Werker & Tees 1983, 2002; Jusczyk et al 1993, 1994; Federici & Wessels 1993; Johnson and Jusczyk, 2001). Second, what experimental evidence we have about the early receptive acquisition of morphology suggests the beginning of such learning occurs somewhere in the second year of life (e.g. Shady, 1996, Santelman and Jusczyk 1998, as well as a brief survey of references in Hayes, 2004).

Of course, this kind of evidence from perception experimentation does not translate directly into a claim about the relative order of acquisition of *productive* grammars or rankings. The entire survey of production data discussed throughout chapter 2 makes it obvious that the production grammar still has much to learn about phonology for several years after the perception revolution of 8-10 months. The anecdotal OO-faith data from section 1.3 also suggests that morphological basehood and its effects on

phonological patterns and paradigms is also still very much under construction throughout early childhood.

7.3.2 The GLA problem with persistent biases and OO-faith

The lag between the acquisition of surface phonotactics and morphological bases provides another kind of winner misparse of the type described in chapter 4. As we've already seen, ERC rows that are added to the Error Cache or Support before any relevant morphological bases have been identified are missing the input structure that allow them to be assigned OO-faith violations.

52) (repeated again from 2)

input	winner ~ loser	Ident[lo]-OO	*ait	* _Λ i	Ident[lo]-IO
(ii)/ɫɔɪrɔɪ/	[ɫɔɪrɔɪ] ~ [ɫɔɪrɔɪ]	e	e	L	W

And so the same argument can be made as in previous sections. The GLA is in danger of end-state overgeneration if the morphological information missing in (52) stays missing long enough – because once this misparsed winner has promoted IO-faith above Markedness, the GLA learner has no way to reverse that ranking.

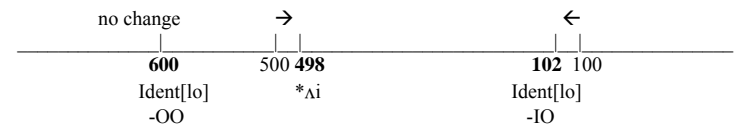
Imagine that we equip the GLA with an *a priori* bias for OO-faith, which ensures that regardless of input data, the OO version of any faithfulness constraint must always remain 20 points above the IO version. Now let us try to learn the distribution of Canadian Raising from section 2. Before morphological discoveries have been made, the learner makes errors that are parsed as in (53) below:

53) Errors during phonotactic learning, with no morphological relations

input	winner ~ loser	Ident[lo]-OO	*ait	* _Λ i	Ident[lo]-IO
(i)/ɫɔɪrɔɪ/	[ɫɔɪrɔɪ] ~ [ɫɔɪrɔɪ]	e	e	L	W

These provide evidence for demoting *_Λi and promoting IO-Ident – and by virtue of this a priori ranking bias, OO-Ident will be promoted as well when IO-Ident gets too close, as shown in (54):

54)a) The GLA learning effect of (51), early in acquisition



54)b) The potential role of the a priori OO-faith bias, later in acquisition



In (54)b), IO-faith has gotten close enough to OO-faith that the a priori bias pushes both constraints up the hierarchy, even though (53) only provides evidence to move the IO version.¹³

¹³ For the reader who wonders how *_Λi can have gotten to a ranking value of 582, considerably higher than its starting value: recall that the grammar could be making other independent errors in which *_Λi preferred the winner. This wouldn't be the case in our Canadian Raising example -- but my real point is not the OO >> IO ranking but rather the reversal of IO-faith and Markedness.

By the later stage ranking above, the GLA has learned the same grammar that BCD initially does – but since it’s stopped making errors, there is no more learning to be done. Fixing a lexical entry for ‘writer’ to include a base won’t cause any errors, because the grammar in 54)b) has high-ranking Faith along *both* dimensions, even though either would do:

55) *No error, post-morphology*

/ɬait + əɬ/	Ident[lo]-OO	Ident[lo]-IO	*ɬi
(a) ɬairəɬ	*!	*!	
(b) ɬairəɬ			*

Kie Zuraw (p.c.) makes the interesting suggestion that the GLA’s general problem with faithfulness and stringency could be addressed by building in a persistent bias for demoting IO-faithfulness independent of errors. For example: the GLA could demote every IO-faithfulness constraints by some small amount every time the grammar is used (or every day, or at some other frequent interval.) This would seem to be the best GLA version of a persistent low-faithfulness bias: it would have the effect of demoting all IO-faith constraints as far as they can go without causing errors – and if OO-faith is now doing the work of preventing errors, IO-faith will be allowed to sink to the bottom. Whether such an approach would provide an adequate answer to all the superset problems raised here is a question open for further investigation.

8. Experimental discussion

8.1 The connection between natural and artificial language learning

The overall results of this experiment matched those that we expect if learners initially rank OO-faith constraints at the top of their grammars. Recall that the learning theory that predicted high-ranking OO-faith was one built in response to superset grammar traps in natural language learning – e.g. McCarthy (1998)’s example in chapter 1 section 3.2.3 of learning static generalizations about non-alternating paradigms. Thus, the fact that these predictions were confirmed here supports the idea that artificial language experiments tap the same kind of phonological knowledge, of constraint rankings and biases, used in natural language acquisition.

From just this study, one might conclude that this connection is only a property of the behaviour of children – that is, that four year olds are still sufficiently engaged in the L1 learning task that their acquisition of artificial forms and paradigms can be influenced by true phonological learning mechanisms. But a number of artificial language learning studies have drawn similar conclusions – even in experiments with adult speakers.

For example, Carpenter (2005) taught native English speakers two patterns of sonority-influenced stress, and found that speakers learning the attested pattern, in which stress is attracted to low vowels, were better at predicting the stress of unfamiliar words than the learners of the opposite pattern of high-vowel stress attraction, which is unattested in the natural language typology. In experiments by Wilson (to appear), English-speaking adults learning patterns of velar palatalization were found to generalize the process from mid vowels to high vowels, but not vice versa, in keeping with the typological fact that natural languages whose velars palatalize next to mid vowels also do

so next to high vowels, but not vice versa. And in Pater and Tessier (2003, 2005), adults were better at learning a phonological alternation that served to meet a static phonotactic generalization of their native language (the English minimal word requirement) than a comparable one that had no such L1 justification.¹⁴ In sum, these results demonstrate that artificial language learning can indeed produce results that accord with a range of assumptions about natural language knowledge and its acquisition.

In the present experiment, one unanswered question is why children are *ever* unfaithful to initial syllable codas in plurals. According to the theory outlined here: once children have learned the suffix ‘del’, the predicted rankings protect base material at the expense of something else (namely affix material or markedness), and therefore a plural noun’s initial syllable coda should remain untouched. This prediction of OO-faithfulness is clearly too strong for my results: section 5 showed that base codas were more faithful than other codas, but that base codas were still much less faithful than onsets in general.

One answer may lie with the mental resources required to implement an OO-faithful grammar: setting up a lexical entry for a closed class affix like ‘plural’, constructing a morphologically-complex input to the phonological grammar and the like. It remains unknown, at least in this methodology, how and when the morphological knowledge that [dəl] is an affix was used online, either to prompt learning via constraint re-ranking or to rule out suboptimal candidates that violated OO-faithfulness. But it seems reasonable to suggest that all of this required a certain amount of concentration and effort, and thus explains part of this variability.

¹⁴ I will also cite the results of Peperkamp and Dupoux (2006) here, but I confess I do not quite understand them yet.

8.2 A potential perceptual confound, and the next step

One alternative reading of this experiment’s results is that the different morphological conditions did not induce different cluster repairs, but rather different percepts. Recall the three morphological contexts in which subjects heard coda segments, e.g. [tʃ]:

- | | | | |
|-----|----------------------------|--------------------------|-----------------------|
| 56) | (a) <i>count singulars</i> | (b) <i>count plurals</i> | (c) <i>mass nouns</i> |
| | [wʌtʃ] | [wʌtʃdəl] | [zɪtʃdm] |

In the mass nouns like 56)c, codas affricates were only ever heard in a cluster, before a following [d]. In the count nouns, however, codas were heard in both the same pre-consonantal context of 56)b, but also word-finally in the related count singular of 56)a. So, it could be that children produced more accurate coda segments in count singulars only because those were the segments that they’d *heard* more accurately. If a subject misheard a coda consonant in the plural form of (56b), their accurate perception of that coda in its related singular could still let them choose the right input form as the base. For the mass noun, however, there is no related form with a word-final version of the coda to suggest an alternative input:

- | | | | |
|--------------|---------------------------------------|----------------------------|--------------------------|
| 57) | <i>Potential perceptual asymmetry</i> | | |
| | <i>count singulars</i> | (b) <i>count plurals</i> | (c) <i>mass nouns</i> |
| (a) sound: | [wʌtʃ] | [wʌtʃdəl] | [zɪtʃdm] |
| (b) subject | | | |
| perceived: | [wʌtʃ] | [wʌtʃdəl], or
[wʌtʃdəl] | [zɪtʃdm], or
[zɪtʃdm] |
| (c) inferred | /wʌtʃ/ | /wʌtʃ + dəl/ | /zɪtʃdm/, or |
| input: | | (from the sing.) | /zɪtʃdm/ |

Knowing precisely how much each participant perceived in each morphological condition is crucial to making claims about the grammars being used or acquired by subjects in the course of experiment.¹⁵ And given its design, this methodology cannot tell us what was perceived.

To eliminate this experimental confound, the best next step is probably to use a similar training methodology, but to test subjects' resulting knowledge using a receptive task -- one that would tap learner's acceptability judgements about *new* forms.

9. Chapter Summary

The experiment reported here provides novel experimental evidence of high-ranking OO-faithfulness constraints in phonological acquisition. When four-year-old children were faced with marked consonant clusters in a novel language, their repairs to those clusters demonstrated preferences for OO-faithfulness over both Markedness and IO faithfulness.

These results provide novel empirical support for the present view of learning, in which learners come to their task with a bias for uniform paradigms independent of data triggers from the target. Further, this experiment also provides novel evidence that children are both willing and able to engage in artificial language learning of this type, particularly in learning new functional material like a plural suffix. Such results may pave the way for a fruitful new brand of experimental work on children's phonologies – because they suggest that artificial language experiments that use more novel materials than traditional wug tests can be used to tap the state of learner's phonological knowledge throughout the course of development.

¹⁵ Thanks to Adam Albright for early discussion of these issues.

BIBLIOGRAPHY

- Alber, Brigit (2001). "Maximizing First Positions," in C. Féry, A. D. Green and R. van de Vijver, eds., *Proceedings of HILP 5*, University of Potsdam.
- Albright, Adam (to appear). "Inflectional paradigms have bases too: evidence from Yiddish." In A. Nevins and A. Bachrach, (eds.) *The Bases of Inflectional Identity*, Oxford: Oxford University Press.
- Albright, Adam and Bruce Hayes (2001). "Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study." *Cognition* 90: 119-161.
- Alderete, John (1999). Head dependence in stress-epenthesis interaction. In Ben Hermans & Marc van Oostendorp (eds.), *The derivational residue in phonological Optimality Theory*, Amsterdam: John Benjamins, pp. 29-50.
- Angluin, Dana (1980). Inductive inference of formal languages from positive data. *Information and Control* vol. 45. pp. 117-135.
- Anttila, Arto (1997). "Deriving variation from grammar". In F. Hinskens, R. van Hout and W. L. Wetzels (eds.) *Variation, Change and Phonological Theory*. Amsterdam: John Benjamins.
- Anttila, Arto (2002). "Morphologically Conditioned Phonological Alternations." *Natural Language and Linguistic Theory* 20.1: 1-42.
- Baertsch, Karen (2002). *An Optimality Theoretic Approach to Syllable Structure: The Split Margin Hierarchy*. Ph.D. dissertation, Indiana University.
- Baker, C. L. (1979). *Syntactic theory and the projection problem*. *Linguistic Inquiry*, 10, 533-81.
- Bakovic, Eric (2000). *Harmony, Dominance and Control*. Ph. D. dissertation, Rutgers University.
- Barlow, Jessica A. (2001). "The structure of /s/ sequences: Evidence from a disordered system." *Journal of Child Language*. 28: 291-324.
- Beckman, Jill N. 1995. Shona height harmony: Markedness and positional identity. In J. Beckman, L. Walsh Dickey, and S. Urbanczyk, (eds.), *Papers in Optimality Theory*. University of Massachusetts Occasional Papers 18. Amherst.: GLSA.pp. 54-75.
- Beckman, Jill N. 1997. Positional faithfulness, positional neutralisation and Shona vowel harmony. *Phonology* 14:1-46.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Beckman, Jill N. 1998. *Positional Faithfulness*. Doctoral dissertation, University of Massachusetts, Amherst.
- Beckman, Jill, Michael Jessen and Catherine Ringen (2006). "Phonetic Variation and Phonological Theory: German Fricative Voicing". In D. Baumer (ed.) *Proceedings of WCCFL25*. Somerville, MA: Cascadilla Press.
- Beckman, Mary and Jan Edwards (2000). "The Ontogeny of Phonological Categories and the Primacy of Lexical Learning in Linguistic Development." *Child Development* 71.1:240-249.
- Becker, Michael (2006). "Learning the lexicon through the grammar - evidence from Turkish". Ms. University of Massachusetts Amherst.
- Benua, Laura. 1997. *Transderivational Identity: Phonological Relations Between Words*. Doctoral dissertation, University of Massachusetts, Amherst. [New York: Garland, 2000.]
- Berko, J. G. (1958) "The child's learning of English morphology," *Word* 14: 150-177.
- Bernhardt, Barbara. H., and Joseph. P. Stemberger (1998) *Handbook of Phonological Development: From the Perspective of constraint-based nonlinear phonology*, San Diego, Academic Press.
- Berwick, Robert (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.
- Blevins, Juliette (1995). "The syllable in phonological theory." In J. Goldsmith (ed.), *The Handbook of Phonological Theory*. Oxford: Blackwell.
- Boersma, Paul (1997). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. Ph.D. dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.
- Boersma, Paul (2001). "Phonology-semantics interaction in OT, and its acquisition." In R. Kirchner, W. Wikeley & J. Pater (eds.) *Papers in Experimental and Theoretical Linguistics*, Vol. 6: 24-35. Edmonton: University of Alberta.
- Boersma, Paul and Diana Appousidou (2003). "Learnability of Latin Stress." In *IFA Proceedings* 25 pp. 101-148. Available as ROA #643.
- Boersma, Paul and Diana Appousidou (2004). "Comparing different Optimality-theoretic learning algorithms fro Latin stress. In V. Chand, A. Kelleher, A. J. Rodriguez, and B. Schmeiser (eds.) *Proceedings of WCCFL*. Somerville, MA: Cascadilla Press. p. 29-42.

- Boersma, Paul and Bruce Hayes (2001). "Empirical tests of the gradual learning algorithm," *Linguistic Inquiry* 32: 45-86.
- Boersma, Paul and Claartje Levelt (2000). "Gradual Constraint-Ranking Learning Algorithm Predicts Acquisition Order". *Proceedings of 30th Child Language Research Forum*, Stanford University. Stanford: CSLI.
- Boersma, Paul and David Weenink (2006) *Praat*. Software package, available at <http://www.fon.hum.uva.nl/praat/>.
- Brame, Michael (1974). "The cycle in phonology: stress in Palestinian, Maltese and Spanish." *Linguistic Inquiry* 5:39-60.
- Broiher, Kevin (1995). Optimality Theoretic Rankings with Tied Constraints: Slavic Relatives, Resumptive Pronouns and Learnability. Ms, MIT. ROA-46.
- Buckley, Eugene. (2003) "Children's unnatural phonology," in Proceedings of the 29th Berkeley Linguistic Society.
- Burzio, Luigi (1998). "Multiple Correspondence". *Lingua*103: 73-109.
- Burzio, Luigi (2000). 'Cycles, Non-Derived-Environment Blocking, and Correspondence.' In J. Dekkers, F. van der Leeuw and J. van de Weijer (eds.), *Optimality Theory: Syntax, Phonology, and Acquisition*. Oxford University Press.
- Carpenter, Angela (2005). BUCLD. "Acquisition of fa Natural vs. an Unnatural Stress System." In A. Burgos, M. R. Clark-Cotton and S. Ha (eds.), *Proceedings of BUCLD29*. Somerville, MA: Cascadilla Press. pp. 134-143.
- Casali, Roderic. (1997). "Vowel elision in hiatus contexts: Which vowel goes?" *Language* 73.3: 493-533.
- Chambers, Jack K. (1973). Canadian Raising. *The Canadian Journal of Linguistics* 18: 113-35.
- Chambers, Kyle, Kristine Onishi and Cynthia Fisher (2003). "Infants learn phonotactic regularities from brief auditory experience." *Cognition* 87: B69-B77.
- Chambless, Della (2006). *Asymmetries in the Acquisition of Consonant Clusters*. Ph.D. dissertation, University of Massachusetts Amherst.
- Cho, Y-M. Y. (1990). *Parameters of Consonantal Assimilation*. Ph.D. dissertation, Stanford University.
- Chomsky, Noam and Morris Halle (1968). *The Sound Pattern of English*. New York: Harper and Row.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Clements, G. N. (1990). "The role of sonority cycle in core syllabification." In J. Kingston & M. Beckman (eds.), *Between the grammar and the physics of speech*. New York: Cambridge University Press.
- Cohn, Abby (1990). *Phonetic and Phonological Rules of Nasalization*. PhD dissertation, UCLA.
- Compton, A.J. and M. Streeter (1977). Child Phonology: Data Collection and Preliminary Analyses. *Papers and Reports on Child Language Development* 7. Stanford, CA: Stanford University.
- Coté, Marie-Hélène (2000). *Consonant cluster phonotactics: A perception-based approach*. Ph.D. dissertation, MIT. Cambridge: MIT Working Papers in Linguistics.
- Curtin, S. and K. Zuraw (2001) "Explaining Constraint Demotion in a Developing System," in B. Skarabela, S. Fish and A. H-J. Do, eds., *Proceedings of the 26th Annual Boston University Conference on Language Development*, Somerville MA, Cascadilla.
- Davidson, L. & M. Goldrick (2003). "Tense, Agreement and Defaults in Child Catalan: An Optimality Theoretic Analysis," In Silvina Montrul, ed., *Selected Papers from the 4th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, Somerville MA, Cascadilla.
- Davidson, L., P. Smolensky & P. Jusczyk (2004) "The Initial and Final States: Theoretical Implications and Experimental Explorations of Richness of the Base," in R. Kager, W. Zonneveld, J. Pater, eds., *Fixing Priorities: Constraints in Phonological Acquisition*, Cambridge, Cambridge University Press.
- Davis, S. (2002) "'Capitalistic' vs. 'Militaristic': The paradigm uniformity effect reconsidered," paper presented at the First North American Phonology Conference, Concordia University, Montreal, Quebec, Canada.
- de Lacy, Paul V. (2002). *The Formal Expression of Markedness*. Doctoral Dissertation, University of Massachusetts, Amherst. Amherst, MA: GLSA.
- de Lacy, Paul V. (2004). "Maximal Words and the Maori Passive". In J. J. McCarthy (ed.) *Optimality Theory in Phonology: A Reader*" Blackwell. pp. 495-512.
- Demuth, K. (1995) "Markedness and the development of prosodic structure," in J. Beckman, ed., *Proceedings of the North East Linguistic Society* 25, Amherst MA, GLSA.
- Demuth, Katherine (1996). Stages in the acquisition of prosodic structure. In E. Clark (ed.), *Proceedings of the 27th Child Language Research Forum*. pp. 39-48. Stanford University: CSLI.

- Demuth, Katherine (2001). "Prosodic constraints on morphological development." In J. Weissenborn & B. Höhle (eds.), *Approaches to Bootstrapping: Phonological, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Amsterdam: John Benjamins. Language Acquisition and Language Disorders Series, vol. 24. pp. 3-21.
- Demuth, Katherine, & E. Jane Fee. (1995). Minimal Words in Early Phonological Development. Ms., Brown University and Dalhousie University.
- Dinnsen, Daniel (1992). "Variation in developing and fully developed phonetic inventories. In C. A. Ferguson, L. Menn and C. Stoel-Gammon (eds.) *Phonological development: Models, research, implications*. Timonium, MD: York Press. pp.191-210.
- Dresher, B. Elan. and Jonathan Kaye (1990) "A Computational Learning Model for Metrical Phonology," *Cognition* 34.2: 137-195.
- Dresher, B. Elan. (1999). "Charting the Learning Path: Cues to Parameter Setting." *Linguistic Inquiry* 30.2: 27-67.
- Elenbaas, Nine (1999). *A unified account of binary and ternary stress: considerations from Sentani and Finnish*. Ph. D. dissertation. Utrecht University
- Elenbaas, Nine and René Kager (1999). "Ternary rhythm and the lapse constraint". *Phonology* 16: 273-329.
- Escudero, Paola and Paul Boersma (2003). "Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm". In S. Arunachalam, E. Kaiser and A. Williams (eds.) *Proceedings of the 25th Annual Penn Linguistics Colloquium (Penn Working Papers in Linguistics 8.1)* pp. 71-85.
- Esper, E. (1925). A Technique for the Experimental Investigation of Associative Interference in Artificial Language Material. *Language Monographs* 1.
- Fennell, Chris T. and Janet F. Werker (2003). 'Early word learners' ability to access phonetic detail in well-known words'. *Language and Speech* 46.2: 245-264.
- Fikkert, Paula (1994). *On the Acquisition of Prosodic Structure*. Ph.D. dissertation, University of Leiden.
- Fikkert, Paula & Clara C. Levelt (to appear). 'How does place fall into place? The lexicon and emergent constraints in the developing phonological grammar'. To appear in P. Avery, B. Elan Dresher & K. Rice (eds.), *Contrast in phonology: Perception and Acquisition*. Berlin: Mouton.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Flack, Kathryn (to appear). 'Templatic morphology and indexed markedness constraints'. To appear in *Linguistic Inquiry* 38.
- Flemming, Edward (1995). *Auditory Representations in Phonology*. Ph.D dissertation, UCLA.
- Flemming, Edward (2001). "Contrast and Perceptual Distinctiveness". In B. Hayes, R. Kirchner and D. Steriade (eds.) *The Phonetic Bases of Markedness*. Cambridge: CUP.
- Fodor, Janet Dean (1998a). "Unambiguous triggers". *Linguistic Inquiry* 29:1-36.
- Fodor, Janet Dean (1998b). "Learning to parse?" *Journal of Psycholinguistic Research*. 27.2.
- Frederici, A. D. and J. E. Wessels (1993) "Phonotactic knowledge of word boundaries and its use in infant speech perception," *Perception and Psychophysics* 54: 287-295.
- Gibson, Edward and Kenneth Wexler (1994). "Triggers". *Linguistic Inquiry* 25.3: p.407-454.
- Gennari, S. and Katherine Demuth (1997). "Syllable omission in Spanish." In E. M. Hughes & A. Green (eds.), *Proceedings of the 21st Annual Boston University Conference on Language Development* vol 1., Somerville, MA: Cascadilla Press. pp. 182-193.
- Gerken, LouAnn (1994). "A metrical template account of children's weak syllable omissions." *Journal of Child Language* 21: 565-584.
- Gnanadesikan, Amahlia. (1995/2004) "Markedness and faithfulness constraints in child phonology," in R. Kager, W. Zonneveld, J. Pater, eds., *Fixing Priorities: Constraints in Phonological Acquisition*, Cambridge, Cambridge University Press.
- Goad, H. and Y. Rose (2004) "Input Elaboration, Head Faithfulness and Evidence for Representation in the Acquisition of Left-edge Clusters in West Germanic," in R. Kager, W. Zonneveld, J. Pater, eds., *Fixing Priorities: Constraints in Phonological Acquisition*, Cambridge, Cambridge University Press.
- Gouskova, Maria (2004). 'Relational hierarchies in OT: the case of Syllable Contact.' *Phonology* 21.2:201-250.
- Grijzenhout, J. and S. Joppen (1998). *First Steps in the Acquisition of German Phonology: A Case Study*. Theory des Lexikons; Arbeit Sonderforschungsbereichs 282, Nr. 110.

- Hale, M. and C. Reiss (1998). "Formal and empirical arguments concerning phonological acquisition," *Linguistic Inquiry* 29: 656-683.
- Hall, Kathleen Currie (2005). "Defining Phonological Rules over Lexical Neighbourhoods: Evidence from Canadian Raising." In J. Alderete, C.-H. Han, and A. Kochetov (eds.) *Proceedings of WCCFL24*. Somerville, MA: Cascadilla Press. pp. 191-199.
- Halle, Morris (1959). *The Sound Pattern of Russian*. The Hague: Mouton.
- Hanson, Kristen and Paul Kiparsky (1996). A parametric theory of poetic meter. *Language* 72: 287-335.
- Hayes, Bruce (1995.) *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.
- Hayes, Bruce (1999). "Phonetically-Driven Phonology: The Role of Optimality Theory and Inductive Grounding". In Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly (eds.), *Functionalism and Formalism in Linguistics, Volume I: General Papers*. John Benjamins, Amsterdam, pp. 243-285.
- Hayes, Bruce (2004a). *OTSoft: Constraint Ranking Software*. Software manual. Available at: <http://www.linguistics.ucla.edu/people/hayes/otsoft/OTSoftManual.pdf>.
- Hayes, Bruce (2004b), "Phonological Acquisition in Optimality Theory: the early stages," in R. Kager, W. Zonneveld, J. Pater, eds., *Fixing Priorities: Constraints in Phonological Acquisition*, Cambridge, Cambridge University Press.
- Hayes, Bruce and Zsuzsa C. Londe (2006). "Stochastic Phonological Knowledge: The Case of Hungarian Vowel Harmony". To appear in *Phonology* 23.1.
- Hayes, Bruce, Bruce Tesar and Kie Kuraw (2003). *OTSoft 2.1*. Software package, available at: <http://www.linguistics.ucla.edu/people/hayes/otsoft/>.
- Inkelas, Sharon, Orhan Orgun and Cheryl Zoll (1997). "The implications of lexical exceptions for the nature of the grammar. In I. Roca (ed.) *Derivations and Constraints in Phonology*. New York: Oxford University Press. pp. 393-418.
- Ingram, David (1989). *First language acquisition: Method, description, and explanation*. Cambridge: CUP.
- Ito, Junko, C. Kitagawa and Armin Mester (1996). Prosodic faithfulness and correspondence: Evidence from a Japanese argot. *Journal of East Asian Linguistics* 5:217-294.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Ito, Junko and Armin Mester (1999) "The structure of the Phonological Lexicon" . In N. Tsujimura, (ed.), *The Handbook of Japanese Linguistics*. Oxford: Blackwell. pp. 62-100.
- Jakobson, Roman and Morris Halle (1956). *Fundamentals of language*. The Hayes: Mouton.
- Jakobson, Roman. (1968) *Child Language, Aphasia and Phonological Universals* (A.R. Kuler, Trans.), Mouton, The Hague. (Originally published as Jakobson, Roman (1941). *Kindersprache, Aphasie, und allgemeine Lautgesetze*. Uppsala: Almqvist & Wiksell.]
- Jesney, Karen (2005). *Chain shift in Phonological Acquisition*. M.A. thesis, University of Calgary.
- Johnson, Elizabeth and Peter W. Jusczyk (2001). "Word Segmentation by 8-Month-Olds: When Speech Cues Count More than Statistics". *Journal of Memory and Language* 44: 548-567.
- Joos, Martin (1942). 'A phonological dilemma in Canadian English.' *Language* 18: 141-144.
- Jusczyk, Peter W., A. Frederici, J. Wessels, V. Svenkenid & A. M. Jusczyk (1993). "Infants sensitivity to the sound patterns of native language words," *Journal of Memory and Language* 32: 402-420.
- Jusczyk, Peter, Paul Luce, and Jan Charles-Luce (1994). "Infants' sensitivity to phonotactic patterns in the native language," *Journal of Memory and Language* 33: 630-645.
- Jusczyk, Peter W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press/Bradford Books.
- Kager, René (1999). "Surface opacity of metrical structure in optimality theory". In B. Herman and M. van Oostendorp (eds.) *The Derivational Residue in Phonological Optimality Theory*. Amsterdam: John Benjamins. pp. 207-245.
- Kager, René (in press). 'Lexical irregularity and the typology of contrast. In K. Hanson and S. Inkelas (eds.) *The Structure of Words*. Cambridge: MIT Press.
- Katz, D. (1987). *Grammar of the Yiddish language*. London: Duckworth.
- Kazazis, Kosta (1969). "Possible evidence for (near)-underlying forms in the speech of a child," in R. Binnick et al, eds., *Proceedings of the Chicago Linguistic Society* 5, Chicago, CLS.

- Keer, Edward (1999). *Geminates, The OCP and the Nature of Con*. Ph.D. dissertation, Rutgers University.
- Kehoe, M. (2000) "Truncation Without Shape Constraints: The Latter Stages of Prosodic Acquisition," *Language Acquisition* 8.1: 23-67.
- Kehoe, Margaret and Geraldine Hilaire-Debove (2004). The Structure of Branching Onsets and Rising Diphthongs: Evidence from the Acquisition of French. In Alejna Brugos, Linnea Micciulla, and Christine E. Smith (eds.), *Proceedings of BUCLD 28*. pp. 282-293. Somerville, MA: Cascadilla Press.
- Kehoe, M. and C. Stoel-Gammon (1997) "Truncation patterns in English-speaking children's word productions," *Journal of Speech, Language, and Hearing Research* 40: 526-541.
- Kenstowicz, Michael (1997). "Base identity and uniform exponence: alternatives to cyclicity." In J. Durand & B. Laks, eds., *Current Trends in Phonology: Models and Methods*. Salford: University of Salford. pp. 363-394
- Kingston, John (1985). *The Phonetics and Phonology of the Timing of Oral and Glottal Events*. Ph.D. dissertation, University of California, Berkeley.
- Kirk, Cecilia and Katherine Demuth (2003). Onset/Coda Asymmetries in the Acquisition of Clusters. In Barbara Beachley, Amanda Brown, and Frances Conlin (eds.), *Proceedings of BUCLD 27*. pp. 437-448. Somerville, MA: Cascadilla Press.
- Kisseberth, Charles (1970). On the functional unity of phonological rules. *Linguistic Inquiry* 1.291-306.
- Levelt, Claartje (1994). *On the Acquisition of Place*. Ph.D. dissertation, Leiden University, HIL Dissertation Series 8.
- Levelt, C. C., N. O. Schiller, et al. (1999) "A developmental grammar for syllable structure in the production of child language," *Brain and Language* 68: 291-299.
- Levelt, C. C. and R. van de Vijver (2004) "Syllable types in cross-linguistic and developmental grammars," in R. Kager, W. Zonneveld, J. Pater, eds., *Fixing Priorities: Constraints in Phonological Acquisition*, Cambridge, Cambridge University Press.
- Lléo, Conxita (1997). 'Filler syllables, Proto-articles and early prosodic constraints in Spanish and German'. In *Language Acquisition: Knowledge, Representation and Processing. Proceedings of GALA 1997*. pp. 251-256.
- Lléo, Conxita (1998). 'Proto-articles in the acquisition of Spanish: Interface between phonology and morphology'. To appear in R. Fabri, A. ortmann and T. parodi

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

(eds.) *Modelle der Flexion: 18. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. Tübingen: Niemeyer.

- Lléo, Conxita (2003). Prosodic licensing of codas in the acquisition of Spanish. *Probus* vol. 15. pp 257-281.
- Lléo, Conxita and Katherine Demuth (1999). 'Prosodic constraints on the emergence of grammatical morphemes: Crosslinguistic evidence from Germanic and Romance languages.' In A. Greenhill, H. Littlefield, & C. Tano (eds.), *Proceedings of the 23rd Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press. pp. 407-418.
- Lléo, Conxita, M. Prinz, Ch. El Mogharbel & A. Maldonado (1996). 'Early phonological acquisition of German and Spanish: A reinterpretation of the continuity issue within the Principles and Parameters model.' In C.E. Johnson and J.H.V. Gilbert (eds.) *Children's Language, Volume 9*. Hillsdale: Lawrence Erlbaum Associates. pp. 11-31.
- Lohuis-Weber, Heleen and Wim Zonneveld (1996). 'Phonological Acquisition and Dutch Word Prosody.' *Language Acquisition* 5.4: 245-284.
- Lombardi, Linda (1991). *Laryngeal features and laryngeal neutralization*. Ph.D. dissertation, University of Massachusetts Amherst. [Garland: 1994].
- Lombardi, Linda (1995). 'Why Place and Voice are different'. In L. Lombardi (ed.) *Segmental Phonology in Optimality Theory: Constraints and Representations*. Cambridge: CUP.
- Lombardi, Linda (1996) "Restrictions on direction of voicing assimilation," *Maryland Working Papers in Linguistics* 4: 89-155.
- Lombardi, Linda (1999). "Positional faithfulness and voicing assimilation in Optimality Theory," *Natural Language and Linguistic Theory* 17.2: 267 – 302.
- Macken, Marlys (1978). "Permitted complexity in phonological development." *Lingua* 44:219-253.
- Macken, Marlys (1980). 'The acquisition of stop systems: A cross-linguistic perspective.' In G. Yeni-Komshian, J. F. Kavanaugh, and C.A. Ferguson (eds.) *Child Phonology Volume 1*. New York: Academic Press. pp. 143-165.
- MacWhinney, B. (1978) "The acquisition of morphophonology," *Monographs of the Society for Research in Child Development* 43.1: whole volume.
- Maiden, Martin (1995). "Vowel systems". In M. Maiden and M. Parry (eds.) *The Dialects of Italy*. London: Routledge. pp. 7-14.

- Maye, Jessica (2000). *Learning Speech Sound Categories From Statistical Information*. Ph.D. dissertation, University of Arizona. Available at: http://www.communication.northwestern.edu/csd/faculty/Jessica_Maye/.
- McCarthy, John J. (1998) "Morpheme structure constraints and paradigm occultation," in M. Catherine Gruber, Derrick Higgins, Kenneth Olson, and Tamra Wysocki, eds., *Proceedings of the Chicago Linguistic Society 5, Vol. II: The Panels*, Chicago, CLS.
- McCarthy, John J. (1999). Sympathy and Phonological Opacity. *Phonology* 16: 331-399.
- McCarthy, John J. (2002). *A Thematic Guide to Optimality Theory*. Cambridge: CUP.
- McCarthy, John J. (2005). "Taking a Free Ride on Morphophonemic Learning". *Catalan Journal of Linguistics* 4: 19-56. [Special issue on morphology in phonology, M.R. Lloret and J. Jiménez, eds.]
- McCarthy, John J. (2006). *Hidden Generalizations: Phonological Opacity in Optimality Theory*. London: Equinox.
- McCarthy, John J. and Alan Prince (1993). *Prosodic Morphology I: Constraint Interaction and Satisfaction*. Technical Report #3, Rutgers University Centre for Cognitive Science.
- McCarthy, John J. and Alan Prince (1995). In J. Beckman, S. Urbanczyk and L. Walsh-Dickey (eds.). 'Faithfulness and Reduplicative Identity.' In *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. Amherst: GLSA. pp. 249-384.
- Mester, Armin and Junko Ito (1989). "Feature predictability and underspecification: Palatal prosody in Japanese mimetics." *Language* 65: 258-93.
- Michelson, Karim (1988). *A comparative study of Lake-Iroquoian accent*. Dordrecht, Boston: Kluwer.
- Mielke, Jeff, Mike Armstrong and Elizabeth Hume (2003). 'Looking through opacity'. *Theoretical Linguistics* 29.1-2: 123-139.
- Morelli, Frida (1999). *The Phonotactics and Phonology of Obstruent Clusters in Optimality Theory*. Ph.D. dissertation, University of Maryland.
- Morrison, Geoffrey S. (2005). "Phonetic naturalness and phonological learnability." Paper presented at the 13th Manchester Phonology Meeting (mfm13).
- Murray, Robert and Theo Venneman (1983). 'Sound Change and syllable structure in Germanic phonology'. *Language* 59: 514-528.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Nagy, N. and W. Reynolds (1995) "Accounting for variable word-final deletion in Optimality Theory," in J. Arnold et al, eds., *Sociolinguistic Variation: Data and Analysis*, Stanford: CLSI.
- Padgett, Jaye (1995). 'Feature Classes'. In J. Beckman, S. Urbanczyk and L. Walsh-Dickey (eds.). 'Faithfulness and Reduplicative Identity.' In *University of Massachusetts Occasional Papers in Linguistics 18: Papers in Optimality Theory*. Amherst: GLSA. pp. 385-420.
- Paradis, Carole and Jean-Francois Prunet (1991). 'Introduction: Asymmetry and visibility in consonant articulations.' In C. Paradis and J.-F. Prunet (eds.) *The special status of coronals: internal and external evidence*. San Diego, CA: Academic Press. pp. 1-28.
- Parker, Steve (2002). *Quantifying the Sonority Hierarchy*. Ph.D. dissertation, University of Massachusetts Amherst.
- Pater, Joe (1997). "Minimal Violation and Phonological Development," *Language Acquisition* 6: 201-253.
- Pater, Joe (2002). "Form and Substance in Phonological Development," in L. Mikkelsen and C. Potts, eds. *Proceedings of the 21st West Coast Conference on Formal Linguistics*, Somerville, MA, Cascadilla Press.
- Pater, Joe (2004a). 'Bridging the gap between perception and production with minimally violable constraints.' In R. Kager, J. Pater, and W. Zonneveld (eds.) *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge: CUP.
- Pater, Joe (2004b). Exceptions in Optimality Theory: Typology and Learnability. Handout from the Conference on Redefining Elicitation: Novel Data in Phonological Theory, NYU. (<http://people.umass.edu/pater/exceptions.pdf>)
- Pater, Joe (to appear). "The Locus of Exceptionality: Morpheme-Specific Phonology as Constraint Indexation". To appear in L. Bateman, M. O'Keefe, E. Reilly and A. Werle (eds.) *UMOP32: Papers in Optimality Theory III*. Amherst: GLSA.
- Pater, Joe and Jessica A. Barlow (2003). Constraint conflict in cluster reduction. *Journal of Child Language*. 30: 487-526.
- Pater, Joe and Andries Coetzee (2005). "Lexically Specific Constraints: Gradience, Learnability and Perception. In *Proceedings of the 3rd Seoul International Conference on Phonology*. Seoul: The Phonology-Morphology Circle of Korea. pp. 85-119.
- Pater, Joe, Chris Stager and Janet Werker (2004) "The Perceptual Acquisition of Phonological Contrasts," *Language* 80.3: 384-402.

- Pater, Joe and Anne-Michelle Tessier (2003). "Phonotactic Knowledge and the Acquisition of Alternations". In M.J. Solé, D. Recasens, and J. Romero (eds.) *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona. pp. 1777-1180.
- Pater, Joe and Adam Werle (2001) "Typology and variation in child consonant harmony," in C. Féry, A. D. Green and R. van de Vijver, eds., *Proceedings of HILP 5*, University of Potsdam.
- Peperkamp, Sharon and Emmanuel Dupoux (2006). 'The Role of Phonetic Naturalness in Phonological Rule Acquisition.' In D. Bamman, T. magnitskaia and C. Zaller (eds.) *Proceedings of BUCLD30*. Somerville, MA: Cascadilla Press. pp. 464-475.
- Pietro Pilar and Maria Bosch-Baliarda (2006). 'The development of codas in Catalan'. In A. Gavarró and C. Lléó (eds.), *Catalan Journal of Linguistics* 5, special issue on L1 Acquisition of Romance.
- Priestly, Tom M. S. (1977). "One idiosyncratic strategy in the acquisition of phonology." *Journal of Child Language*, 4, 45-66.
- Prince, Alan (1997). "Paninian relations," paper presented at LSA Summer Institute, Cornell University.
- Prince, Alan (2002a). "Arguing Optimality," in A. Coetzee, A. Carpenter and P. V. de Lacy, eds., *Papers in Optimality Theory II*, Amherst, GLSA.
- Prince, Alan (2002b). "Entailed Ranking Arguments". Manuscript, Rutgers University. ROA-500.
- Prince, Alan (2005). Lecture notes from LSA Summer Institute course, MIT/Harvard.
- Prince, A. and P. Smolensky (1993/2004), *Optimality Theory: Constraint interaction in generative grammar*, Oxford, Blackwell.
- Prince, A. and B. Tesar (2004) "Learning Phonotactic Distributions," in R. Kager, W. Zonneveld, J. Pater, eds., *Fixing Priorities: Constraints in Phonological Acquisition*, Cambridge, Cambridge University Press.
- Pulleyblank, Douglas & William J. Turkel (1998). "The logical problem of language acquisition in optimality theory," in P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis & D. Pesetsky, eds., *Is the best good enough? Optimality and competition in syntax*. Cambridge, MA: MIT Press, 399-420.
- Pulleyblank, Douglas & William J. Turkel (2000) "Learning Phonology: Genetic Algorithms and Yoruba Tongue-Root Harmony," in Joost Dekkers, Frank van der

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Leeuw, & Jeroen van de Weijer, eds., *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford: Oxford University Press, 554-591.
- Pycha, Anne, Pawel Nowak, Eurie Shin and Ryan Shosted (2003). "Phonological Rule-Learning and Its Implications for a Theory of Vowel Harmony." In G. Garding and M. Tsujimura (eds.) *Proceedings of WCCFL22*. Somerville, MA: Cascadilla Press. pp. 423-435.
- Revithiadou, A. and Marina Tzakosta (2004) "Markedness Hierarchies vs. Positional Faithfulness and the Role of Multiple Grammars in the Acquisition of Greek," *Proceedings of Generative Approaches to Language Acquisition (GALA) 2003*, Utrecht, LOT Occasional Series.
- Rice, Keren, and Peter Avery (1989). "On the interaction between sonorancy and voicing." *Toronto Working Papers in Linguistics* 10: 65-82.
- Rice, Keren, and Peter Avery (1991). "On the relationship between coronality and laterality." In C. Paradis and J.-F. Prunet (eds.) *The special status of coronals. Internal and external evidence. (Phonetics and Phonology, vol. 2)*. San Diego: Academic Press.
- Roark, Brian and Katherine Demuth (2000). Prosodic Constraints and the Learner's Environment: a Corpus Study. In S. C. Howell, S. A. Fish, and T. Keith-Lucas (eds.), *Proceedings of BUCLD 24*. Somerville, MA: Cascadilla Press. pp. 597-608.
- Robins, R. H. (1957). 'Vowel nasality in Sundanese'. *Studies in Linguistic Analysis*. London: Basil Blackwell. pp. 87-103.
- Rose, Yvan. (2000) *Headedness and Prosodic Licensing in the LI Acquisition of Phonology*. Ph.D. dissertation, McGill University.
- Saffran, Jenny, Richard Aslin and Elissa Newport (1996). "Statistical learning by 8-month-old infants." *Science* 274: 1926-1928.
- Saffran, Jenny, Elissa Newport and Richard Aslin (1996). "Word Segmentation: The Role of Dsistributonal Cues". *Journal of Memory and Language* 35: 601-621.
- Saladis, Joanna and Jacqueline S. Johnson (1997). 'The production of minimal words: A longitudinal case study of phonological development.' *Language Acquisition* 6: 1-36.
- Santelmann, Lynn and Peter W. Jusczyk (1998). "Sensitivity to Discontinuous Dependencies in Language Learners: Evidence for Limitations in Processing Space," *Cognition* 69: 105-134.

- Selkirk, Elisabeth O. 1994. Optimality Theory and featural phenomena. Lecture notes, LING 730, University of Massachusetts, Amherst.
- Shady, Michele E. (1996) *Infants' sensitivity to function morphemes*. Doctoral dissertation, SUNY Buffalo, Buffalo, NY.
- Smith, Jennifer L. (1999). "Theories of Faithfulness: Implications for Learnability." Handout from talk at RumJClam Workshop, Rutgers University.
- Smith, Jennifer L. (2000) "Positional faithfulness and learnability in Optimality Theory," in Rebecca Daly and A. Rehl, eds., *Proceedings of ESCOL99*, Ithaca, CLC Publications.
- Smith, Jennifer L. (2001). "Lexical Category and Phonological Contrast". In R. Kirchner, J. Pater and W. Wikely (eds.) *PETL6: Proceedings of the Workshop on the Lexicon in Phonetics and Phonology*. Edmonton: University of Alberta. pp. 61-72.
- Smith, Jennifer L. (2002). *Phonological Augmentation in Prominent Positions*. Ph. D. dissertation, UMass Amherst.
- Smith, Neilson V. (1973). *The Acquisition of Phonology: A Case Study*. Cambridge, MA: Cambridge University Press.
- Smolensky, Paul (1996a). The initial state and 'Richness of the Base.' Technical Report JHUCogSci-96-4. ROA-154.
- Smolensky, Paul (1996b). "On the comprehension/production dilemma in child language," *Linguistic Inquiry* 27: 720-731.
- Smolensky, Paul and Geraldine Legendre (2006). *The Harmonic Mind*. Cambridge: MIT Press.
- Stager, Christine L. and Janet F. Werker (1997). 'Infants listen for more phonetic detail in speech perception than in word-learning tasks.' *Nature* 388: 381-2.
- Stampe, David (1969). The acquisition of phonetic representation. In *Papers from the Fifth Regional Meeting, CLS*. pp 443-454. Chicago: Chicago Linguistic Society.
- Steriade, Donca (1995). "Positional Neutralization." Ms, UCLA.
- Steriade, Donca (1997). "Phonetics in Phonology: the Case of Laryngeal Neutralization". In M. Gordon (ed.) *UCLA Working Papers in Phonology* vol. 2.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Steriade, Donca (1999). "Alternatives to the syllabic interpretation of consonantal phonotactics," in O.Fujimura B.Joseph and B.Palek (eds.) *Proceedings of the 1998 Linguistics and Phonetics Conference*, The Karolinum Press, 205-242
- Steriade, Donca (2000). "Paradigm Uniformity and the Phonetics-Phonology Boundary," in M. Broe, and J. Pierrehumbert, eds., *Papers in Laboratory Phonology 5*, Cambridge, Cambridge University Press.
- Stites, J., K. Demuth & C. Kirk. (2004) "Markedness versus frequency effects in coda acquisition," in , eds., *Proceedings of the 28th Annual Boston University Conference on Language Development*, Somerville MA, Cascadilla.
- Strujke, Caro (2002). *Existential Faithfulness: A Study of Reduplicative TETU, Feature Movement and Dissimilation*. Outstanding Dissertation in Linguistics. NY: Routledge.
- Tesar, Bruce (1995). *Computational Optimality Theory*. Doctoral Dissertation, University of Colorado at Boulder.
- Tesar, Bruce (1997). An iterative strategy for learning metrical stress in Optimality Theory. In E. Hughes, M. Hughes and A. Greenhill (eds.) *Proceedings of the Twenty-First Annual Boston University Conference on Language Acquisition*. Somerville, MA: Cascadilla Press. pp. 615-626.
- Tesar, Bruce (1998). Using the mutual inconsistency of structural descriptions to overcome ambiguity in language learning. In EDS, *Proceedings of the Twenty-Eighth Conference of the North Eastern Linguistic Society*. 469-483.
- Tesar, Bruce (2000). Using inconsistency detection to overcome structural ambiguity in language learning. Technical Report RuCCS-TR-58, Rutgers Center for Cognitive Science, Rutgers University. ROA-426. pp 1-41.
- Tesar, Bruce, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani and Alan Prince (2003). "Surgery in language learning". In G. Garding and M. Tsujimura (eds), *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, Somerville, MA, Cascadilla Press. pp. 477-490.
- Tesar, Bruce and Paul Smolensky (1998). "Learnability in Optimality Theory." *Linguistic Inquiry* 29. pp. 229-268.
- Tesar, Bruce and Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
- Tessier, Anne-Michelle (2006). "Learning Stringency Relations and the Structure of Faithfulness". 80th LSA meeting, Albuquerque, New Mexico.

- Trubetskoy, N (1939). *Principles of Phonology*. C. Baltaxes, (trans.) Berkeley/Los Angeles: University of California Press.
- Tzakosta, Marina (2004). *Multiple parallel grammars in the acquisition of stress in Greek L1*. Ph. D. dissertation, Leiden University.
- Ussishkin, Adam (2000). *The Emergence of Fixed Prosody*. Ph.D. dissertation, University of California, Santa Cruz.
- Van der Pas, Brigit (2004). Contiguity in phonological acquisition. In J. van Kampen & S. Baauw (eds.) *Proceedings of GALA 2003* (vol. 2, pp. 353-364). Utrecht: LOT Occasional Series 3.
- Velleman, Shelley and Marilyn Vihman (2000). 'The construction of a first phonology.' *Phonetica* 57: 255-266.
- Velleman, Shelley and Marilyn Vihman (2002a). 'Whole-word Phonology and Templates: Trap, Bootstrap or Some of Each?' *Language, Speech and Hearing Services in Schools* 33:9-23.
- Velleman, Shelley and Marilyn Vihman (2002b). 'The Optimal Initial State'. Manuscript, University of Massachusetts and University of Bangor.
- Walker, Rachel (1998). *Nasalization, Neutral Segments and Opacity Effects*. Ph.D. dissertation, UC Santa Cruz.
- Werker, Janet F., L. B. Cohen, V. L. Lloyd, M. Casasola and Chris L. Stager (1998) "Acquisition of Word-Object Associations by 14-Month-Old Infants" *Developmental Psychology* 34.6: 1289-1309.
- Werker, Janet F., & Suzanne Curtin (2005) "PRIMIR: A developmental model of speech processing," *Language Learning and Development* 1.2: 197-234.
- Werker, Janet F. and Richard Tees (1983). 'Developmental changes across childhood in the perception of non-native speech sounds.' *Canadian Journal of Psychology* 37:278-286.
- Werker, Janet F. and Richard Tees (2002). 'Cross-language speech perception: Evidence for perceptual reorganization during the first year of life.' *Infant Behaviour and Development* 25:121-133.
- Wexler, K. and R. Manzini (1987) "Parameters and learnability in binding theory," in T. Roeper and E. Williams, eds., *Parameter Setting*, Dordrecht, Reidel.
- Wilson, Colin (2000). *Targeted constraints: An approach to contextual neutralization in Optimality Theory*. Ph.D. dissertation, Johns Hopkins University.

Tessier, Anne-Michelle (2007). *Biases and Stages in Phonological Acquisition*. Ph.D. dissertation, UMass Amherst

- Wilson, Colin (2001). 'Consonant cluster neutralisation and targeted constraints'. *Phonology* 18: 147-197.
- Wilson, Colin (2003). "Experimental Investigation of Phonological Naturalness". In G. Garding and M. Tsujimura (eds.) *Proceedings of WCCFL22*. Somerville, MA: Cascadilla Press. pp. 533-546.
- Wilson, Colin (to appear). "Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization." To appear in *Cognitive Science*.
- Winslow, Nicholas (2003). *Incorporating Exceptions in Optimality Theoretic Learnability*. Honors Thesis, University of Massachusetts, Amherst.
- Wittgott, Mary M. (1982). *Segmental evidence for phonological constituents*. Ph.D. dissertation, University of Texas, Austin.
- Wolf, Matthew (2005). An autosegmental theory of quirky mutations. In John Alderete, Chung-hye Han, and Alexei Kochetov (eds.), *Proceedings of the 24th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, pp. 370-378.
- Zoll, Cheryl (1996). *Parsing below the segment in a constraint-based framework*. Ph.D. dissertation, UC Berkeley.
- Zoll, Cheryl (1998). *Positional Asymmetries and Licensing*. Ms., MIT. Available at <http://roa.rutgers.edu/files/282-0998/roa-282-zoll-4.pdf>.
- Zuraw, Kie (2000). *Patterned Exceptions in Phonology*. Ph.D. dissertation, UCLA.