Learning in the Face of Infidelity:
Evaluating the Robust Interpretive Parsing/Constraint Demotion Model of Optimality Theory Language Acquisition

Joshua Ian Tauberer

A THESIS

in

Linguistics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Master of Arts

2008

Supervisor of Thesis

Graduate Group Chairman

Contents

Abstract

Error Driven Constraint Demotion with Robust Interpretive Parsing (RIP/CD; Tesar and Smolensky 2000) is a model of language learning for Optimality Theory. There is still much left to be understood about its relationship to the subset problem and its usefulness for learning phonological alternations. We evaluate RIP/CD using simulations, with an approach new to the OT literature in several ways: 1) it is based on a Markov model to capture the complete behavior of the learning system (following Niyogi and Berwick 1996), 2) we look across the complete typology of target languages in a domain involving faithfulness violations, and 3) we consider different criteria for successful learning.

The primary type of OT analysis considered in the learnability literature has been metrification, but the problems that arise for RIP/CD are quite different for systems of segmental alternation, e.g. when there are faithfulness violations. We discuss these problems formally and evaluate their impact in two simulations involving constraints related to oral and nasal vowel alternations, the English plural morpheme's voicing, and denasalization of consonants.

Our two major findings are as follows. First, the problem of the final refinement step in EDCD raised by Boersma (2008b) persisted in RIP/CD, and a 'Variationist RIP/CD' gives the learner a better chance of learning languages. In fact, such a learner is guaranteed to eventually converge on a grammar that at least corresponds to a superset of the target language. The second major finding is that constraining Universal Grammar, meaning restricting the set of the learner's possible initial states, does not seem to solve the subset problem at all. No markedness-over-faithfulness ranking, for instance, allowed all target languages to be learnable.

This thesis provides a foundation for moving forward with OT learnability. We provide baseline numbers to compare future models against, to judge whether they face the same set of problems and how well they face these problems.

Parts of this thesis were presented at the 32nd Annual Penn Linguistics Colloquium, February 22-24, 2008 and those parts will be published in the conference proceedings, the Penn Working Papers in Linguistics Volume 15.1 (2009).

I would like to thank my master's supervisor Charles Yang for his direction, feedback, and honest perspective of the field, my second reader Gene Buckley for his comments and sharing his expertise in phonology, Catherine Lai for an unprintable expletive that helped in the condensation of my PLC talk, and the Phonetics Lab's Splunch group and the audience and reviewers of the 32nd Penn Linguistics Colloquium for their feedback.

Mystery
All my life has been a mystery
You and I were never ever meant to be
It's why I call my love for you a mystery

Different country
You and I have always lived in a different country
And I know that airline tickets don't grow on a tree
So what kept us apart is plain for me to see

That much at least is not really a mystery

Estuary
I live in a houseboat on an estuary
Which is handy for my work with the Thames water authority
But I know you would have found it insanitary

Insanitary

Taken a violent dislike to me
I'd be foolish to ignore the possibility
But if we ever actually met you might have hated me
Still, that's not the only problem that I can see

Dead since 1973
You've been dead now... wait a minute, let me see
Fifteen years come next January
As a human being you are history

So why do I still long for you
Why is my love so strong for you
Why did I write this song for you
Well, I guess it's just a mystery

— *Hugh Laurie, 'Mystery' (1987)*

Learning in the Face of Infidelity: Evaluating the Robust Interpretive Parsing/Constraint Demotion Model of Optimality Theory Language Acquisition

1 Introduction

Computational models or algorithms for language acquisition tell us a great deal both about the faculty of language as well as about our theories of grammar. The 'constraint demotion' (CD) family of algorithms pursued by Bruce Tesar and Paul Smolensky and later others has provided a useful foundation for exploring learnability in Optimality Theory (OT). In CD algorithms, the learner reranks constraints over time, by demoting them, until his grammar is compatible with what he hears in his environment. Tesar and Smolensky's contribution to language acquisition with CD was finding a means for a learner to use parsed structure to infer something about the target grammar. A model based on CD lets us ask questions such as whether the whole typology of languages predicted by orderable constraints should all be expected to be attested (Boersma 2003, Alderete 2008), why particular longitudinal patterns in child development occur (Demuth 1995), and why universal rankings involving e.g. positional faithfulness might arise from imperfect learning (Boersma 2008a). The most well-understood model of language acquisition in the CD family is Error Driven Constraint Demotion with Robust Interpretive Parsing (RIP/CD) (Tesar and Smolensky 1996, Tesar 1998a, Tesar and Smolensky 2000), which is the focus of this thesis.

Tesar (1998a:132) has been quite optimistic about the core of the algorithm, the 'constraint demotion procedure':

> [Paul Smolensky and I (Tesar) have] presented the constraint demotion procedure, which completely solves the subproblem of the relationship between full structural descriptions and grammars. . . . As a solution to this subproblem, constraint demotion has several appealing properties: (a) it applies to all lin-

> guistic analyses within the OT framework, not just those of, say, stress, or even phonology; (b) it is guaranteed to find a correct ranking; and (c) it is quite fast.

We first must unpack what Tesar is talking about. He is not making claims about RIP/CD, the full model of language learning, but instead to EDCD, a simplified case where the learner hears input–output pairs from his environment directly, rather than just outputs as in RIP/CD. For instance, in RIP/CD the learner hears an overt form such as 'opacity' and must make an inference to undo velar softening to an arrive at the underlying form 'opaque+ity'. Or, he hears the overt metrical pattern ˈHL˛HL and must infer the locations of parsed feet as in (ˈHL)(H)L. This inferential step could go wrong. In the quotation above, Tesar refers to the easier case where the learner hears an input–output pair and hidden structure, what he calls a full structural description, directly, and does not need to make an inferential step. Nevertheless, despite addressing a simpler problem, Tesar's excitement was premature. Guaranteed to find a correct ranking it most certainly is not. Despite the claims in Tesar and Smolensky (1996, 2000), which apply to the simpler EDCD problem only, Boersma (2008b) presented a EDCD simulation showing OT languages (though not linguistically motivated languages) that the EDCD learning procedure did not find a correct ranking for.

What we know is that *some* learning algorithm must exist to guarantee at least probably approximately correct (PAC) learning of any OT language given input–output pairings as in EDCD. Vapnik-Chervonenkis dimension is a measure of the complexity of a grammatical system, such as OT, in terms of the capacity of the system to create a large typology of languages. Riggle's (to appear) computation of the VC-dimension of OT as finite (linear in the number of constraints in the system), provided an upper-bound on the difficulty of learning a constraint ranking. Finite VC-dimension guarantees PAC learnability, that there is some algorithm that with high probability hypothesizes a language that is not too far off from the correct language. (See Stabler 2008 for background on formal learnability theory as it relates to linguistics.)

If the input–output pairings are known, the learning problem is essentially solved. Boersma (2008b) proposed a revision to EDCD to fix the failure he found, creating what he called Variationist EDCD. If Boersma is correct that this revised algorithm is guaranteed to converge to a correct grammar, this would be good progress — although still limited to the easier EDCD problem.

But, 'the relationship between full structural descriptions and grammars' (Tesar 1998a:132) is just a small piece of the language learning problem. Learners do not have access to underlying forms or full structural descriptions. Just imagine what kind of benefit this would have for syntax if the learner could see the underlying form, the structure, of sentences: the learner could set various parameters merely by inspecting the tree! Though initial forays into syntactic acquisition, e.g. in Wexler (1978), did make the simplifying assumption of access to hidden structure, more recent work (Gibson and Wexler 1994, Niyogi and Berwick 1996, Fodor 1998) along with RIP/CD assume this is too much to expect of the learner.

The work on syntactic acquisition and that on OT acquisition have diverged in their responses to this problem. The syntactic line of work, e.g. Gibson and Wexler's (1994) Triggering Learning Algorithm, side-stepped the issue of not knowing underlying forms. The learner seeks a grammar that can merely parse the input, and does not use the parsed structure to learn something about the grammar.

CD algorithms rely crucially on parsed structure. In RIP/CD the learner faces the even harder problem of inferring underlying forms and full structural descriptions from what he hears. Robust Interpretive Parsing is a step added into the EDCD model to have the learner guess hidden levels of structure. So while Constraint Demotion has proven itself as a viable means to update an OT hypothesis, based on the work on EDCD and Variationist EDCD, the question of the viability of Robust Interpretive Parsing remains.

Models of language acquisition are evaluated from several angles: whether a child is expected to be able to carry out the algorithm given human limits of memory and computation, whether given any input from the idealized environment the algorithm will make a

best guess as to what the grammar is, how often this best guess is correct or close enough, and how long it takes and how many examples from the environment the algorithm needs to reach it. EDCD and RIP/CD succeed on the first evaluation: They are indeed psychologically plausible. However, while Variationist EDCD seems to be guaranteed to make the correct guess, the convergence proofs of Tesar and Smolensky (1996, 2000) do not apply to RIP/CD because either the inputs or the full structural descriptions are not known to the learner. Examples exist where convergence does not occur: Tesar (1998a:141) himself noted that a learner employing RIP/CD to learn metrical stress can get stuck alternating between two incorrect grammars. After hearing one word from the environment the learner demotes the TROCHAIC constraint below the IAMBIC constraint, but then on another input the learner does the reverse, cycling indefinitely. Apoussidou and Boersma 2004 confirmed this in simulations. It has been well known, then, that RIP/CD is not guaranteed to find a ranking at all, let alone a correct one.

And what about a correct one? RIP/CD faces several problems, some noted in literature, others not. The success of RIP/CD for OT systems of metrical phonology have been considered on several occasions (Tesar and Smolensky 1996, Tesar 1998a, Tesar and Smolensky 2000, Apoussidou and Boersma 2004), and in systems involving faithfulness violations but in learning algorithms besides EDCD and RIP/CD (e.g. Boersma and Hayes 2001, Hayes 2001, Alderete and Tesar 2002, Prince and Tesar 2004), but few times has there been an analysis, and no simulations, of OT systems involving faithfulness violations in EDCD or RIP/CD (Smith 2000, Boersma 2003). The trend of results is that EDCD and RIP/CD are no silver bullet. We ask in this thesis 'how bad is it?', specifically for learning plausible natural languages involving input–output faithfulness violations.

We evaluate the robustness of RIP/CD below. Our approach to simulating a RIP/CD learner is new to OT language acquisition research in that our simulation:

- is based on a Markov model to capture the complete behavior of a learner, in the spirit of Niyogi and Berwick (1996), rather than a simulation that randomly selects a

particular order of presentation of stimuli to the learner.

- involves an OT system admitting faithfulness violations.
- considers different notions of 'success' in learning to measure the degree to which the algorithm works out-right and also after it is granted some leeway with regard to the subset problem.
- is run first the RIP/CD algorithm described by Tesar and Smolensky and then run with an update to RIP/CD following Boersma's (2008b) modifications to EDCD in Variationist EDCD.

In addition we:

- look across the complete factorial typology of languages predicted by the OT system and see how many of those can be learned by a RIP/CD learner.
- consider several conceptions of Universal Grammar (UG), meaning different initial states for the learner.

The thesis begins by reviewing the RIP/CD procedure and discussing why we have opted to evaluate RIP/CD rather than newer CD algorithms which have appeared to supersede RIP/CD in the literature (section 2). In section 3 we discuss, at an abstract level, limitations to the RIP/CD procedure, including those reported previously and several new limitations presented here. Sections 4 and 5 describe the methodology and results of our simulations. Section 6 concludes.

2 Review of RIP/CD

2.1 Overview

RIP/CD has its roots in Gold (1967) and Wexler (1978), where learners were proposed to maintain at any given time a single hypothesis of what the grammar is that is generating the language he hears in this environment. With each new token from the language the learner hears (a 'positive example' of an element from the target language), the learner may change his hypothesis, but never necessarily knowing whether he has ever finally arrived at the *right* hypothesis. Gold's 'identification in the limit' is when the learner arrives at a hypothesis that he will never further revise, although the learner isn't aware that he has reached this state. How the learner updates his hypothesis will depend on the grammatical system involved. (This model does not include the use of 'negative evidence', explicit indication that a token is *not* in the target language.)

When grammars are reducible to a finite number of universal parameter switches, as in the Principles and Parameters syntactic tradition, the learner may flip switches after hearing a sentence from his environment, hopefully flipping the right ones. In the Triggering Learning Algorithm (TLA; Gibson and Wexler 1994, Niyogi and Berwick 1996), the learner changes his hypothesis when he encounters a a sentence that he cannot parse with his current grammar. He then chooses a parameter at random and flips its value. If the resulting grammar can parse the token, he keeps the new parameter value, otherwise he flips the parameter back and waits for another token. It is impossible to know which parameters were responsible for the failure of the grammar to parse the token — it might be a conspiracy of parameter settings, for instance.

Tesar and Smolensky improved on these essentially random movements between hypotheses because in OT one can see which constraints are at play in the analysis of a linguistic object, and this is informative about what changes to the hypothesis grammar may be worthwhile. The contributions of CD and RIP/CD to models of language acquisition

include:

- Using the learner's current hypothesis grammar to parse input *and* infer the input's hidden structure.
- Generating (possibly faulty) implicit negative evidence and updating the grammar to rule out this evidence.
- Using (possibly faulty) knowledge of hidden structure to update the learner's hypothesis. (This part is similar to Fodor's (1998) structural triggers learner who parses with a supergrammar to determine parameter values that must be set in order to parse an input sentence in any UG-allowed grammar.)

RIP/CD, like TLA, is a psychologically plausible algorithm in that it is memory-less, meaning the learner only retains in memory a hypothesis grammar and not, for instance, a bag of all the words he has ever heard. It is also computationally easy, in that it does not require the learner to do much work after each token (word). A more specialized learning algorithm specific for OT is both a blessing and a curse, however: since the algorithm will not be applicable to other components of language or cognition generally, two distinct learning algorithms may be needed (Charles Yang p.c.).

RIP/CD is based on the EDCD algorithm, which addresses the simpler problem discussed in the introduction. EDCD will always settle on a final hypothesis before N^2 'useful' tokens, where N is the number of constraints in the system. This is quite fast, and remarkable, considering the learner is faced with a space of $N!$ grammars to choose from. TLA, on the other hand, is not guaranteed to settle at all, and it has a much smaller space of just 2^N grammars (where N here is the number of parameters). Unfortunately, this guarantee does not apply to RIP/CD.

At a high level, RIP/CD works by generating implicit negative evidence. "Positive evidence" is the information given to the learner asserting what is *in* the language. "Negative

evidence" is the information asserting what is *not* in the language, for example an adult correcting the learner by saying a form was wrong. It is commonly assumed, for the purposes of learnability research, that the learner is not provided negative evidence. One of RIP/CD's contributions is a means for the learner to (imperfectly) generate negative evidence. When the learner updates his hypothesis, he reorders the constraints so his generated negative evidence is ruled out and the actually observed positive evidence is ruled in. The reordering follows the procedure of Constraint Demotion. This process then repeats.

Robust Interpretive Parsing, the double-edged sword that allows the learner to confront the lack of knowledge of hidden structure but at the price of sometimes making a bad guess, is primarily responsible for the *positive* evidence — that is, the hidden structure of the positive evidence. Constraint Demotion, on the other hand, is responsible for generating the *negative* evidence and updating the grammar.

Several assumptions are made about the learner and his environment in the OT learnability literature that are adopted here:

- The adult grammar is a fully stratified OT grammar. Phenomena such as variation in outputs and gradience in grammatically are not considered.
- The learner's linguistic environment comprises outputs from this grammar alone, and contains no mistakes (noise).
- The learner's linguistic environment comprises surface phonological strings. Phonetic forms and ambiguity at levels besides phonology are not considered.

RIP/CD is described well elsewhere (Tesar and Smolensky 1996, 2000), so we will merely review the process by way of an example in the following sections.

2.2 Initial Procedure

For the example we will use constraints related to the agreement in nasality between vowels and the coda in English (Kager 1999:31), which we use for the simulations later on. The

three constraints we will use in this section are:

- *$\tilde{\text{V}}$: No nasal vowels.
- *V_{ORAL}N: No oral vowels before a nasal consonant in the coda.
- Ident(nas): Nasality faithfulness (for vowels only — consonants cannot change nasality here).

plus Max and Dep implicitly always undominated.

Let us place a learner in the environment of English, with adults ranking the constraints as *V_{ORAL}N $\gg$ *$\tilde{\text{V}}$ $\gg$ Ident(nas). Under these constraints, vowel nasality always corresponds with the nasality of the following consonant — there are therefore no vowel nasality contrasts in the language.

The first step in RIP/CD is to assign the learner an initial state, which is to say give him an initial hypothesis grammar. RIP/CD is silent on what the learner's first hypothesis should be, as it is an empirical question which initial states are going to put the learner off on the right foot. There may be (and are) initial states from which the learner cannot learn certain languages. In the simulations below, we say that UG specifies a certain set of grammars as permissible initial states, and the learner draws one grammar from that set at random at the beginning of his learning process. Let us give our learner the grammar *$\tilde{\text{V}}$ $\gg$ Ident(nas) $\gg$ *V_{ORAL}N, which admits in outputs only oral vowels, regardless of a following nasal consonant.

2.3 Robust Interpretive Parsing

The rest of RIP/CD is a repeating process. An adult first chooses something to say, a base form B, and generates from it the optimal candidate C and its corresponding overt form in the language, a token or overt form O.

Let us first start with an aside about analyses of metrical phonology. The base form in a metrification analysis is something like HLHLH, a sequence of heavy and light syllables.

Candidates include the locations of parsed feed, as in (HL)(HL)H. But output tokens do not have that information, only the location of primary stress ˈHLˌHLH (e.g. Tesar and Smolensky 2000). That is, we allow the learner to 'hear' without the possibility of mistake 1) the number of syllables in the utterance, 2) each syllable's weight, and 3) which syllables have primary and secondary stress. Crucially, we do not give the learner the parsing of feet. He must infer that.

In the case of segmental phonology, base forms, candidates, and outputs are all concatenations of segments. A base form /oupɛɪk-ɪti/ yields the surface form /oupæs-ɪti/. In our model of language acquisition for segmental phonology, we make the simplifying assumption of cutting out phonetics. We assume the adult communicates the surface phonological form /oupæs-ɪti/ directly to the learner. In other words, the optimal candidate is, unlike in metrical analysis, identical to the output token.

Let's now say an adult wants to emit the lexical entry /pæ̃nt/. He then runs the tableau to find the optimal candidate for this base and generate an output:

pæ̃nt	*V_{ORAL}N	*$\tilde{V}$	IDENT(nas)
pænt	*!		*
☞pæ̃nt		*	

Figure 1: An adult English speaker wants to emit /pæ̃nt/. The optimal candidate is /pæ̃nt/, which is what is emitted.

In this case, the base form B, optimal candidate C, and overt token O are all the same, /pæ̃nt/. The reader can imagine that the optimal candidate may have come out differently. The learner next hears this token O. Actually, how the adult generated O is not important, since the learner does not have access to that information. We go through the process above here only to guarantee that the tokens the learner receives are consistent with some particular adult grammar.

The next step of RIP/CD is for the learner to guess the base form B' and the full

structural description or candidate C' that the adult used to generate the token. This is the process of Robust Interpretive Parsing (Tesar and Smolensky 1996, 2000, Tesar 1998b).

Here again there is a difference between segmental phonology and the types of metrical phonology analyses that have been considered in the CD literature (cited on page 4) which did not permit faithfulness violations. When a learner 'hears' an output such as ˈHL,HLH, he does not know the optimal candidate in the adult's grammar (the location of feet), as we discussed, but he does immediately know the base form that generated the candidate, HLHLH, since syllables were not permitted to be added or removed, or their weight changed. Thus he determines B' immediately. If the grammatical system allows faithfulness violations, e.g. in the type of segmental phonology analysis of concern in this thesis, the learner does not immediately know the base forms of outputs he hears. Alternations, i.e. faithfulness violations, may have occurred. On the other hand, he *can* see the optimal candidate directly from our simplifying assumption about cutting out phonetics. Thus, he determines C' instead immediately.

So how can a learner know the underlying form of a phoneme that undergoes alternation without already knowing the cause of the alternation? (And likewise for the location of feet in analyses of metrification.) Of course, if the learner could do this step accurately he would have already pretty much solved the problem of language acquisition. This is the bootstrapping problem of language acquisition. RIP/CD confronted this problem in a new way by making use of the hypothesis grammar that the learner maintains: the learner hopes that his hypothesis grammar is close enough to the adult grammar that he can use it for parsing and not be too far off.

Robust Interpretive Parsing is a function that guesses from an output token O and a grammar H the base form B' and candidate C' that underly the output (Tesar and Smolensky 1996:10). The base form and candidate pair yielded by Robust Interpretive Parsing satisfies two conditions: 1) C''s corresponding output is indeed O, and 2) compared to all other such pairs of base forms and candidates, it is the most harmonic under H. This

is a similar procedure to Lexicon Optimization proposed by Prince and Smolensky (1993), the main difference that in Robust Interpretive Parsing C' need not be the optimal output for B' under H.

Our learner, who has just heard O = /pæ̃nt/ and maintains a hypothesis grammar G, now performs Robust Interpretive Parsing. He is looking for B' and C'. C' he knows: it is also /pæ̃nt/ because it is the only candidate that matches what he heard. Two base forms (or in general many) could generate that candidate: /pænt/ and /pæ̃nt/. He selects as B' the one that is optimal when we consider a special tableau like the one in Figure 2. Rows in this tableau represent 'base → candidate' pairs. We assess markedness violations on the candidate halves of the pairs, and faithfulness violations by comparing the candidate in each row to its corresponding base form in that same row. Note how the markedness violations are the same in each row: because the candidate halves C' has already been determined.

	*Ṽ	IDENT(nas)	*V_{ORAL}N
pænt → pæ̃nt	*	*!	
☞pæ̃nt → pæ̃nt	*		

Figure 2: Robust Interpretive Parsing of the output /pæ̃nt/.

The grammar is the learner's current hypothesis. The learner here is only looking for the best base form, and this means the base form in the row with the fewest faithfulness violations. Here he chooses /pæ̃nt/, which is indeed B, the base form the adult had in mind. The learner gets it right.

2.4 Constraint Demotion

The learner next runs the base form B' that he computed through Robust Interpretive Parsing through a tableau according to his current hypothesis grammar. This will yield an optimal candidate *for him*, which we will label as N. See Figure 3.

The learner concludes that N = /pænt/ is optimal (for him). Now, the learner believes,

pæ̃nt	*Ṽ	IDENT(nas)	*V$_{\text{ORAL}}$N
☞pænt		*	*
pæ̃nt	*!		

Figure 3: The learner runs the base form B' through a tableau according to his current hypothesis grammar and concludes that /pænt/ is optimal (for him).

according to Robust Interpretive Parsing, that in the adult grammar B' = /pæ̃nt/ makes the candidate C' = /pæ̃nt/ optimal. If N, what is optimal for B' under the learner's current grammar, is not the same as C', the learner can conclude he has the wrong grammar, and he will update his grammar using Constraint Demotion (CD), described next, so that the adult candidate C' comes out optimal instead.

Tesar calls N the 'loser' because it is what we want to not be optimal for the learner once CD has been performed, and C' the 'winner' because it is what ought to win if the learner is to have the right grammar. Though we have no particular desire to arbitrarily alter terminology, we agree with Boersma (2003) that these terms are confusing because the loser is for a short while an optimal candidate. Instead, we will call C' the positive evidence or positive candidate, because it is what the learner thinks is *in* the language and N, if it differs from C', as the implicit negative evidence or negative candidate, since it is what the learner concludes is *not* in the language. (We also refrain from using the term 'input', which Tesar and Smolensky have used to mean underlying form as in the OT tradition, but elsewhere is the data the learner receives. We use 'base' or 'underlying' form in the former case and 'token' in the later case.)

CD is the core of RIP/CD. It is a process of reordering constraints based on 1) a filled-in tableau and 2) candidates marked as the positive and negative candidates. The learner wants to make his grammar look more like the target grammar, although the target grammar is of course not available to him. CD specifies a procedure to attempt to do so: Demote constraints (in the hypothesis grammar) that the positive but not the negative candidate

violated so that they are all in the stratum immediately below the highest constraint that the negative but not the positive candidate violated. This will help because once the constraints are moved around in this way, the positive candidate will be more harmonic than the negative candidate. (In the case of constraints that assess multiple violations on a candidate, one looks and whether the positive or negative candidate violated it more times.)

In the example, the positive candidate /pæ̃nt/ uniquely violated *Ṽ, while the negative candidate /pænt/ uniquely violated IDENT(nas) and *V_{ORAL}N. CD then says to move the first constraint into the stratum immediately below IDENT(nas). That yields a new grammar shown in the tableau below, which also shows that the positive example is now optimal. Note that *Ṽ and *V_{ORAL}N are now not mutually ranked: they are in the same stratum.

pæ̃nt	IDENT(nas)	*V_{ORAL}N	*Ṽ
pænt	*!	*	
☞pæ̃nt			*

Figure 4: After Constraint Demotion, the learner has this grammar, in which the positive candidate comes out optimal.

No action is taken by the learner when a token is compatible with the currently hypothesized grammar, that is, when the negative and positive candidates are the same, or if there are no constraints uniquely violated by the positive and negative candidates.

The procedure above repeats until the learner no longer makes any changes to his hypothesis grammar.

Grammars are stratified hierarchies, and tableaux here are evaluated during both Robust Interpretive Parsing and in generating the negative candidate by pooling the violations of all of the constraints in a stratum (as in Tesar and Smolensky 1996, 2000 but not Tesar 1998b).

2.5 The Last Step: Refinement

While Tesar considered adult grammars to always be drawn from the set of fully stratified grammars, CD can create a non-fully stratified grammar, as in the example here. It is another problem how a learner might turn his final hypothesis grammar from the last CD step, if it is not fully stratified, into a fully stratified grammar. Tesar and Smolensky (2000:49) suggested, "At the end-point of learning, the hierarchy may not be fully ranked: the result is a stratified hierarchy with the property that any further refinement into a fully ranked hierarchy will correctly account for all the learning data," where 'refinement', defined later, meant choosing any grammar that preserves all of the domination relations. For instance, the grammar {A,B} $\gg$ {C,D} has four refinements: A $\gg$ B $\gg$ C $\gg$ D, B $\gg$ A $\gg$ C $\gg$ D, A $\gg$ B $\gg$ D $\gg$ C, B $\gg$ A $\gg$ D $\gg$ C. Boersma (2008b) unfortunately determined that this claim was patently false, showing that the learner can choose a hypothesis (in the limit) for which only some refinements are compatible with the target language. We found this to be the case in our simulations.

Because the choice of refinement matters, we include this step in our model. But since Tesar and Smolensky have not indicated how the learner chooses a particular refinement, we assume the learner chooses one at random.

Boersma (2008b) proposed a solution to this problem. He suggested modifying the EDCD algorithm so that the refinement procedure is used during the constraint demotion iterations, besides also at the end of the EDCD process. When the learner receives an input–output pair, he makes a temporary refinement and computes the negative candidate N by running the input form through a tableau under this refinement. Then he forgets the refinement and continues with constraint demotion on the original stratified grammar. Boersma called this Variationist EDCD, and includes a proof of its guaranteed convergence to any target grammar. Though, he does not verify the proof with a simulation.

We extend RIP/CD here in a similar way, to create Variationist RIP/CD. In Variationist RIP/CD, the learner creates a refinement of his current hypothesis grammar before the

Robust Interpretive Parsing step. He computes the base form and full structural description of the output token using the refinement, and then as with Variationist EDCD computes the negative candidate using the refinement but continues with constraint demotion on the original hypothesis grammar. The simulations below were run with both the standard RIP/CD and Variationist RIP/CD to determine whether Boersma's fix to EDCD fixes problems with RIP/CD as well.

2.6 Comparison of Algorithms for OT Grammar Learning

So far, we have seen four OT learning algorithms: EDCD, RIP/CD, Variationist EDCD, and Variationist RIP/CD. Their differences can be summarized by the table in Figure 5. The constraint demotion and final refinement steps are the same for all four algorithms.

	(standard)		Variationist	
	EDCD	RIP/CD	EDCD	RIP/CD
Parsing Structure of the Input	Learner is given parsed structure.	Robust Interpretive Parsing	Learner is given parsed structure.	Robust Interpretive Parsing on 'refined' grammar
Generating a Negative Candidate	Learner evaluates a tableau		Learner creates a temporary refined grammar and then evaluates a tableau	

Figure 5: Comparison of OT learning algorithms.

We next briefly touch on other algorithms in the greater constraint demotion family in this section and explain why it is too early to end the research program on RIP/CD.

2.6.1 (Multi-)Recursive Constraint Demotion

RIP/CD has effectively been superseded in the literature by, roughly, two sets of algorithms: batched forms of constraint demotion and Multi-Recursive Constraint Demotion (MRCD; Tesar 1997).

Recursive Constraint Demotion (RCD; Tesar and Smolensky 1996, 2000) is one of several

'batched' versions of constraint demotion in which the learner makes use of *all* of the (unique) tokens he has observed in the past to make a new hypothesis. The learner must maintain tokens in memory, a step away from the memory-less nature of RIP/CD. This increased reliance on memory — one that is clearly psychologically implausible at face value — allows RCD to be robust to a set of inputs that is not consistent with a single grammar. In that case, RCD will terminate, unlike RIP/CD, signaling to the learner the problem (who could then take some more enlightened move). Of course, a psychologically plausible version of RCD could be made, say by requiring the learner to remember only the last 100 tokens, but it is an open question of just how well RCD would work with imperfect or incomplete memory. The algorithm does not improve on RIP/CD if we hold to the idealization that the learner receives consistent input, though it serves as the basis for other algorithms.

Biased Constraint Demotion (BCD; Prince and Tesar 2004) was proposed as an extension to RCD to bias analyses in a particular direction in order to address the subset problem. Low Faithfulness Constraint Demotion (LFCD; Hayes 2001) is a batched version of CD with similar goals.

With regard to BCD, the authors are quite explicit that RCD and BCD are to be taken not as learning algorithms on their own but instead are intended as a component of an online learning procedure, such as MRCD (Prince and Tesar 2004:11). When BCD is used within MRCD, it is called Biased Multi-Recursive Constraint Demotion (BMRCD).

MRCD, the second major alternative to RIP/CD, carries out RCD or another equivalent batched process over multiple and possibly very many hypotheses simultaneously. Because RCD identifies inconsistent input, it can be used to prune away incorrect hypotheses. MRCD has serious questions of psychological plausibility — in terms of the usual time and memory factors. The algorithm, like RCD, requires the learner to maintain more information in memory than a single hypothesis. Rather, the learner maintains a set of plausible hypotheses, adding to and pruning the set as tokens are processed.

When the learner encounters a token, he considers every possible parse of that word. In the case of metrical stress, the number of parses of any word is fairly limited — there are only so many ways to parse feet. But when any base form could underlie any surface form, when there is the possibility of faithfulness (MAX, DEP, IDENT-IO, etc.) violations, considering all possible parses one at a time may itself be intractable. It is likely for this reason that the authors of MRCD explicitly limit the domain of their problem to cases where the base form is trivially recognizable, i.e. cases similar to metrical stress (Tesar 1997:7). Furthermore, each possible parse may multiply the number of hypotheses the learner must maintain in memory, leading to a second possible unbounded growth. It remains to be seen whether in linguistically plausible situations these potential unbounded growths actually occur, but it is not the focus here.

We stay with Tesar and do not consider MRCD applicable where the underlying forms of what the learner hears are not transparent, and so we do not consider MRCD further. We recognize that it is, however, an important subject of further study even for phonemic alternation learning.

2.6.2 Gradual Learning Algorithm

The Gradual Learning Algorithm (GLA; Boersma 1997, Boersma and Hayes 2001) is a stochastic approach to grammar learning. As opposed to the approaches mentioned so far, GLA treats a constraint's rank as a probability distribution over a continuous scale. Constraints are not necessarily ranked at discrete positions either above, below, or at the same level as other constraints, as in standard OT. Instead, their positions are probabilistic, and this accounts for variability in output. The remainder of GLA is, however, almost entirely parallel to EDCD. The learner compares what he heard to the output that is optimal according to his hypothesis grammar and then repositions those constraints with asymmetric violations. The primary difference in GLA is that in repositioning constraints, ranking adjustments are small so that the learning is gradual and robust to noisy input.

(See Boersma and Hayes 2001, Pater 2008 for a summary of convergence results.) Like EDCD and RIP/CD, GLA is psychologically plausible.

Also as with EDCD, GLA assumes the learner has access to the underlying form of each adult form heard. We thus expect that the same problems described below for RIP/CD would apply to 'RIP/GLA' (Boersma 2003) as well, that is, GLA extended with Robust Interpretive Parsing so that the learner could infer underlying forms.

On the other hand, Boersma (2003) and Apoussidou and Boersma (2004) have demonstrated that learners that fail with EDCD and RIP/CD can succeed with GLA and RIP/GLA. There are, therefore, substantive differences to the two models of learning, and GLA appears to have a promising future.

3 Theoretical Limitations

3.1 When Robust Interpretive Parsing Goes Wrong

Robust Interpretive Parsing is a fragile process. Tesar and Smolensky (2000) describe two cases where it can lead to the wrong analysis, which we sketch here, and we describe two more cases later in this section.

When Robust Interpretive Parsing fails for analyses of metrification, it will do so because it chooses the wrong candidate C'. The base form can be observed by the learner directly, so he has no trouble deciding on B'. The first case that Tesar and Smolensky described as foiling this process is when an interpretation (C') is chosen that 'cannot possibly be optimal', one that cannot be an optimal output in any language (any fully stratified grammar). They wrote that this may occur when parsing a 'structural description that contains both iambic and trochaic feet' (63), with the result that the RIP/CD procedure cycles between two grammars indefinitely. They do not discuss whether this parsing error necessarily leads to a cycle, however.

Their second problematic case is when the candidate selected by Robust Interpretive Parsing is 'harmonically bound' by the candidate generated by the learner's current hypothesis grammar. In this case, Robust Interpretive Parsing not only chooses a candidate C' that is not optimal under any language, but further has all of the violations that the negative candidate N has plus additional violations. This is problematic because the constraint demotion step crucially compares only the violations that each candidate has but not the other. The violations of C' but not N determine which constraints to demote, while the violations of N but not C' determine where to demote them to. Since N in this case has no unique violations, constraint demotion cannot update the grammar.

Considering OT for segmental phonology, the positive candidate can be seen directly so Robust Interpretive Parsing cannot make these types of errors. On the other hand, if faithfulness violations are possible then Robust Interpretive Parsing might choose the wrong

base form.

The first problem we find is Robust Interpretive Parsing choosing a base form that *could* underly the output, but not the one the adult had in mind. In cases where contrasts in the base are neutralized, two base forms yield the same output. A RIP/CD learner would have no way of telling which of the two base forms is responsible for any given occurrence of one of these outputs. In an alternate situation where the adult intended to convey the underlying form /pænt/, this time a non-nasal vowel, it will still be realized as /pæ̃nt/ because agreement in nasality outranks preserving the nasal contrast. When the learner performs Robust Interpretive Parsing, he will get the same result as last time, /pæ̃nt/ as the underlying form. In fact, the learner will always choose the base form that is closest to the output because it will have the fewest violations. Here, /pæ̃nt/ is a correct parse, just not the correct parse that the adult had in mind, so the consequences for the learner is that he may miss some evidence — i.e. how should the base form /pænt/ be realized? This will leave the learner free to make any hypothesis regarding /pænt/, and he may choose a grammar that gets it wrong (while still getting /pæ̃nt/ correct). The richness of the base principle in OT (Prince and Smolensky 1993:191) indicates that this is a problem. According to richness of the base, language-specific facts cannot be derived by placing restrictions on the set of possible underlying forms, which are instead conceived to be universal. A grammar is correct only if it generates a valid output for every potential lexical entry. What happens to /pænt/ is important, regardless of whether it actually underlies any observed adult forms.

If the only I–O constraint is IDENTITY, then this may be the worst problem facing RIP/CD (although it has consequences discussed in section 3.2). No one, however, has explored what kind of problems Robust Interpretive Parsing may have if I-O constraints are allowed that differ from IDENTITY. Such constraints may have unintended consequences for RIP/CD. We can imagine a completely unmotivated constraint such as the following:

> C_1: Assess one violation for the input–output correspondence b—x, two for a—x, and three for a—y (and no violations for b—y).

There's nothing about the nature of Optimality Theory that prevents one from using such a constraint, though it may have no value to linguistics. Nevertheless, let us look at what happens if there are two base forms a and b, two candidates x and y, and just this one constraint. In the adult grammar, a will be realized as x and b as y, shown in Figure 6.

a	C_1
☞x	**
y	***!

b	C_1
x	*!
☞y	

Figure 6: Computing the language of a toy grammar.

However, when the learner attempts Robust Interpretive Parsing on x, he will choose the wrong base form, even though he has the correct adult grammar already (there being no choice for how to rank a single constraint). The tableaux for Robust Interpretive Parsing of x and y are shown in Figure 7.

Parsing x:

	C_1
$a \rightarrow x$	**!
☞$b \rightarrow x$	*

Parsing y:

	C_1
$a \rightarrow y$	*!**
☞$b \rightarrow y$	

Figure 7: Robust Interpretive Parsing of x and y in the toy grammar.

Parsing incorrectly when the learner already has the correct adult grammar is a dire situation for the learner. If there were other constraints in the grammar, the learner would likely at this point change his grammar to something *other than* the adult grammar, taking a step in the wrong direction.

C_1 is perhaps completely unmotivated for linguistic phenomena, but there has not been a serious analysis yet of what happens with other types of correspondence constraints: contiguity, linearity, and even maximality and dependence, let alone two-level well-formedness constraints (Kager 1999:378), output-output correspondence constraints, and sympathy. What kinds of constraints play nice with RIP/CD and what kinds of constraints break

Tesar's system?

All that being said, we put this last issue aside for the remainder of this thesis as it does not arise in the simulations.

3.2 The Subset Problem

Research into learning algorithms within OT have shifted away from RIP/CD primarily because of RIP/CD's inability to escape a general problem for learning algorithms called the subset problem. The subset problem occurs when the learner's grammar produces all of the overt forms in the target language (the extension of the target grammar), but also produces additional forms. The problem is that if the learner receives only positive evidence, as in RIP/CD or EDCD, the learner gets no explicit instruction that the additional forms that he *thinks* are grammatical actually are not. In this case, the target language is a subset of the learner's current hypothesis. If the learner were to compare his language with what he has heard so far in his environment, he would merely think he has been very unlucky in having not heard various words yet. The Subset Principles defined variously in the literature going back to 1985 (see Fodor and Sakas 2005 for a summary) are approaches to the subset problem by constraining the learning procedure, and possibly also the initial state of the learner, so that narrower hypotheses are considered before their supersets.

The subset problem in OT was recognized at the origins of constraint demotion. Tesar and Smolensky (1996) addressed the subset problem explicitly, especially as it related to the relationship between faithfulness and markedness constraints. When markedness constraints like NoCoda are ranked over faithfulness constraints, a restricted e.g. CV-syllable-only language is generated. Ranking faithfulness higher than markedness means that more syllable types are available by preserving contrasts present in the base; it yields a superset of the first language. If the learner hypothesizes the grammar with more forms, the one with faithfulness high, Tesar and Smolensky asked how the learner could deduce that that the more narrow grammar, one that excludes CVC syllables, is the correct grammar. Lacking

negative evidence, explicit instruction that some forms are ungrammatical, the learner has to be a little creative to keep his grammar as narrow as possible. The presence of grammatical alternations would, they claimed, provide evidence of the high ranking of markedness constraints. Additionally, attributing the following to Alan Prince, initial hypotheses could be required to rank faithfulness constraints lowest. Thus in the absence of evidence *for* CVC syllables that leads to demoting markedness constraints below faithfulness constraints, the learner would start in and be left with the most restrictive grammar. Tesar and Smolensky favored this approach because it "accords well" (34) with child acquisition data (see Demuth 1995, Gnanadesikan 1995, Levelt et al. 2000). Smith (2000) looked specifically at the ramifications in RIP/CD of the subset problem caused by analyses of positional faithfulness implemented as position-specific faithfulness constraints: these constraints must start off in the initial state ranked below markedness constraints but above general faithfulness constraints.[1] We return to why this must be so momentarily.

It is interesting that this result is in direct conflict with Hale and Reiss's (2003) strongly-worded conclusion for the 'logical necessity' of learners starting with grammars that make all possible contrasts. Hale and Reiss are right to say that 'Without access to a difference in representation, the phonetic difference between the two vowels cannot be evaluated' (236). They argue that if a learner employs a language that lacks a certain contrast — say, the oral/nasal vowel distinction — the learner has no way of creating the contrast because he cannot linguistically distinguish tokens from the two categories, precisely because he lacks the contrast. As a result, they claim, learners must start off with a hypothesis that the language being learned makes all possible contrasts, losing some of the contrasts over time as appropriate to their linguistic environment. They even quote Zenon Pylyshyn to reinforce the structure of their argument, ' "[i]f you believe P, and you believe that P entails Q, then ... Q may be true." ' Hale and Reiss conflated the ability for a learner to represent a

[1] If EDCD is the term for the simpler problem described in our introduction, Smith created a new adaptation of EDCD which happens to be equivalent to RIP/CD.

contrast he hears with the inclusion of the contrast in the learner's hypothesis grammar, and this difference is exploited in CD models of language acquisition. As we've seen, a learner may be able to linguistically represent the oral/nasal contrast in base forms (in fact, following richness of the base this is a necessity) while nevertheless hypothesizing a grammar in which markedness constraints neutralize the contrast overtly. Tesar and Smolensky have shown, then, that a learner can both posses a rich representational system from the start but nevertheless exhibit a growing inventory of outputs, rather than shrinking. Hale and Reiss's 'necessity' is, rather, impossible. We thus would like to counter-quote from Star Trek's Spock, or Sherlock Holmes: 'Whenever you have eliminated the impossible, whatever remains, however improbable, must be true.'

Neither Tesar and Smolensky (1996) nor another head-on attack of the subset problem in Smolensky (1996) adequately explained why it is in constraint demotion that initially highly ranked faithfulness constraints (but not markedness constraints) lead to the subset problem. It is true that highly ranked faithfulness constraints produce languages that are supersets of the target language. However, it is *not* true that learners employing RIP/CD necessarily get stuck whenever they posit a superset-language hypothesis. We considered above a contrived example where the learner infers the wrong base form of a token during the process of Robust Interpretive Parsing, even though the learner had hypothesized exactly the right grammar. Inferring an incorrect input–output mapping, the learner would be expected to make *some* change to his grammar, and this would cause him to change his hypothesis from the correct grammar — which is a superset of itself — to some other grammar. At least when there are I–O constraints beyond simple IDENTITY, it is possible for the learner to not be stopped by the subset problem. (In those cases the learner might have bigger problems to worry about, however.)

IDENTITY constraints pose a bigger problem than just the subset problem for RIP/CD, however. A conspiracy prevents faithfulness constraints from *ever* being demoted in RIP/CD[2].

[2]This observation seems to go back to Smolensky (1996:14), though he does not mention Robust Interpretive Parsing in explaining the problem.

As we noted earlier, Robust Interpretive Parsing always posits a base form that has the fewest faithfulness violations compared to the output. Given an output, say /pæ̃nt/, if an identical form exists as a base form, then Robust Interpretive Parsing will conclude that that was the base form, even if other base forms such as /pænt/ might be the underlying form of the output token that the adult has in mind. As a result, the learner will never posit a base form against which the output incurs a faithfulness violation, and without ever encountering a faithfulness violation, the faithfulness constraints will never be operated on by the Constraint Demotion step. They will never be demoted, although other constraints may be demoted over them. If they start ranked high, the learner has no hope to learn a language where they are lower ranked.

Much additional work on the subset problem in OT has focused on algorithms in the CD family tailored specifically to addressing the problem: Biased (Multi-Recursive) Constraint Demotion (Alderete and Tesar 2002, Prince and Tesar 2004, McCarthy 2005) and Low Faithfulness Constraint Demotion (Hayes 2001). The successes of these improvements can only be measured against the capabilities of prior models, such as RIP/CD, however, which is part of the reason why we believe the research agenda of this thesis to be important.

3.3 Ramifications for Universal Grammar

Several types of RIP/CD's failures have been pointed out in the literature: Robust Interpretive Parsing giving forms that cannot possibly be optimal, incorrectness of the final 'refinement' step, and the subset problem. We have pointed out or clarified several other related problems: the failure to hypothesize all base forms, hypothesizing incorrect base forms when there are unusual constraints, and the fact that IDENTITY constraints are never demoted. The research program among most of the authors that have considered these problems has been to hold out hope that even if RIP/CD does not work in general, that is, when the learner is free to start with any initial state (his initial grammar), perhaps there are particular initial states from which all languages are learnable with RIP/CD.

Smolensky (1996) for instance noted that in order to get around the problem of the inability to demote faithfulness constraints, they must start off in the initial state ranked below all markedness constraints, so that markedness constraints can be demoted below them. Smith (2000) suggested further that positional faithfulness constraints must outrank other faithfulness constraints in the initial state. One must wonder then how much specification of UG is necessary to get RIP/CD off the ground, and is this the extent of the problem? We suggested above the other types of input–output correspondence constraints may have unexpected effects on Robust Interpretive Parsing and, so, also on what might need to be specified about UG.

If two faithfulness constraints are freely ranked typologically (in attested languages), however, then it is impossible to imagine a UG from which both orders could be learnable. If the position of faithfulness constraints cannot be changed, then their position relative to other faithfulness constraints will be fixed according to the learner's initial state. It is easy to devise an OT system that exhibits this problem:

- The base forms and candidates are xx, xy, yx, and yy.
- The constraints in the system are Ident[Initial] for faithfulness in initial position, Ident[Final] for faithfulness in final position, and OCP, the obligatory contour principle, assessing a violation for xx and yy.

There are six grammars in the typology given by these three constraints. The input–output relations are given in Figure 8. The grammars come in pairs if we group them by their input–output relations. In the second pair, the Ident[Initial] constraint always outranks the Ident[Final] constraint, and in the third pair the reverse. If we expect the learner to arrive at the correct input–output relation, then there is no initial state that will let the learner arrive at a grammar from all three groups, since for at least one group the order of the identity constraints would need to change. On the other hand, if we only expect the learner to arrive at the right overt language, the right set of overt forms, then the situation

Grammar↓ Base Form→	xx	xy	yx	yy
Id[I] ≫ Id[F] ≫ OCP	xx	xy	yx	yy
Id[F] ≫ Id[I] ≫ OCP	xx	xy	yx	yy
Id[I] ≫ OCP ≫ Id[F]	xy	xy	yx	yx
OCP ≫ Id[I] ≫ Id[F]	xy	xy	yx	yx
OCP ≫ Id[F] ≫ Id[I]	yx	xy	yx	xy
Id[F] ≫ OCP ≫ Id[I]	yx	xy	yx	xy

Figure 8: Six grammars and their input–output relations.

is different. There are two, rather than three, groupings: the first two grammars yield { xx, xy, yx, yy } while the latter four yield { xy, yx }. For each of these languages, a learner can find one grammar that yields it with the identity constraints in any order, whichever order is compatible with the learner's initial state. The ramifications for UG depend on what we expect of the learner.

But even in a domain without faithfulness violations, the choice of initial state has a large effect on what is learnable. Tesar and Smolensky (2000) reported simulation results for a 12-constraint system explaining foot parsing in metrical stress. Constraints included Parse, Iambic, All-Feet-Left, etc. Three particular initial grammars were tested, and the number of languages out of 124 chosen to be tested that were learnable were reported. By learnable, they meant that the RIP/CD algorithm setttled on a particular grammar. The unlearnable ones presumably then caused the algorithm to indefinitely cycle between options. The first initial state was a grammar comprised of a single stratum containing all of the constraints unranked. With this initial state, 60 (48%) of the languages were learnable. The second initial grammar comprised two strata, the first containing the two foot form constraints and the second containing the remaining constraints. In this grammar, 76 (61%) of the languages were learnable. Starting with the last grammar, a three-stratum grammar (the details of which do not concern us here), a learner would learn 97 (78%) of the languages. Tesar (1998a) reported for a similar simulation a 'designed' initial state starting from which a learner could learn all 104 out of 104 languages tested. It is very difficult to

interpret these results. First, Tesar and Smolensky are curiously silent on whether when the learner settled on a grammar if the learner in fact settled on a grammar equivalent in some way to the target garmmar, or if the learner went astray and got stuck. Second, one would like to know whether the languages that were not learnable are in fact attested. It would be a success if the remaining 27 languages under the third initial grammar turned out to all be unattested.

When evaluating a CD learning procedure, there are simply too many degrees of freedom. If a language is not learnable, we say that explains why it is not attested. If it *is* attested, we say that we must have gotten the wrong initial state. If we can't find an initial state that works, we may blame it on the subset problem. The only way to get a handle on just what RIP/CD has accomplished is to simulate all of these cases and see exactly where RIP/CD works and exactly where it doesn't.

4 Simulating RIP/CD with a Markov Model

4.1 Building a Markov Model for Language Learning

This paper presents the results of several simulations of RIP/CD. The simulations of the type reported in Tesar and Smolensky (2000), for instance, subject the learner to a particular (but random) order of surface forms. The results are then not directly generalizable to learners who receive the tokens in other orders; the effect of the random choices made during the simulation is unclear and unreported. Following Niyogi and Berwick (1996), a Markov model captures the complete behavior of a learner exposed to a distribution of surface forms by essentially computing all possible paths the learner may take in response to all possible orders of surface forms. The results of a Markov model, i.e. the computed probabilities of succesful learning, are exact.

A Markov model is described by a state transition diagram, such as the one shown in Figure 9. Here, each of the states (cricles) in the model represents a hypothesis grammar that the learner has at any given time. On hearing a surface form from the environment, the learner moves from one state into another or back into the same state. Transitions in the Markov model (arrows) represent an action after hearing a token from the environment, changing the hypothesis grammar. The probability of each transition (indicated next to the arrow) is determined by the probability distribution of tokens in the learner's environment and the result of one step of the RIP/CD algorithm on each of the tokens.

The simulations were performed by constructing a large matrix of transition probabilities and multiplying the matrix out to determine the exact probability that the learner would be in any state after hearing a certain number of randomly drawn (with replacement) examples from his environment. There is nothing stochastic in this procedure. The simulations were run using a new Python program written by the author. (EDCD simulations are already possible using the Praat program thanks to the work of Paul Boersma (Boersma and Weenink 2008). Praat's simulations are stochastic rather than based on the Markov model

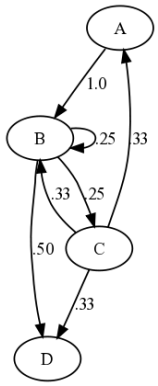

Figure 9: A Markov model.

method used here.) Further details of the simulations not indicated below are presented in Appendix A.

Both the standard RIP/CD procedure and a modified version along the lines of Boersma's (2008b) Variationist EDCD were used, which we call Variationist RIP/CD. Variationist EDCD was intended to correct the problem of refinement. Boersma's modification to EDCD (which, as the reader may recall from earlier, addresses the simpler problem when the learner has direct access to underlying forms) was to use the 'refinement' procedure earlier in the algorithm. (See section 2.5.) Because there may be a large number of refinements for any grammar, and too many to consider in a simulation of this sort, we sampled a small number of refinements at each refinement step. In this case only, the Markov model approximates the true transition probabilities.

We considered as target languages all of the languages that arise from all permutations of the constraints involved in the simulation (which are listed in section 5; there were 54 languages in the second simulation). The probability distribution of the surface forms in the learner's environment was computed for a particular target language at a time. Because of the large number of target languages, data collection would have been expensive and indeed impossible for those permutations that don't correspond to an attested human language.

Based loosely on the richness of the base principle in OT, which says roughly that any base form could be a lexical entry in any language (Prince and Smolensky 1993:191), we assumed that all possible base forms are in fact lexical entries in every language, each base form being used by the adult population with equal frequency. The probability distribution over overt forms that are given to the learner are computed by running each base form through a tableau one-by-one and noting the resulting output (automatedly of course).

4.2 What Constitutes Successful Learning?

For a learner to have learned anything, he must arrive at a hypothesis which no future token could cause him to abandon (Gold 1967). Such a hypothesis may be compatible with all of the overt forms of the language in which the learner is immersed (in fact it might be the correct hypothesis), but it may also be a hypothesis that is clearly wrong but which RIP/CD cannot get the learner out of. Recall that RIP/CD does not update the hypothesis if either the positive or negative candidates do not uniquely violate a constraint. If this occurs, the learner could get prematurely stuck. When the learner gets stuck, we say the learner has halted.

When has a halted learner successfully learned (i.e. converged on) the target language?

We might idealize the learning environment as consisting of tokens from a source that employs a particular grammar, i.e. a particular order of constraints. In that case, we might call a learner successful if he converges on that very same order of constraints. However, such a notion of success would be unrealistically narrow. Language users cannot see the inner workings of a tableau, and it is common for different orders of constraints to be equivalent in their mappings from base forms to winning candidates. As a result, learners could not possibly distinguish which particular order of constraints an adult employs among a set of equivalent orders.

A second and less strict notion of success is if the learner arrives at a grammar whose 'extension' (i.e. its yield or outputs) is the same the target's extension — that is, if the

grammars have the same surface forms, though the surface forms may correspond to different underlying forms. Learners cannot see underlying forms, and the RIP/CD algorithm does not directly address the problem of learning underlying forms, especially in the case of alternations and opacity (see Alderete and Tesar 2002 and McCarthy 2005). Extensional success then seems to be an appropriate criterion to use. This is related to the concept of weak generative capacity.

However, even 'extensional success' may not be fair to RIP/CD. We know RIP/CD was never designed to address the subset problem (see the citations on page 26 for those that were). Tesar and Smolensky may have envisioned a separate process of, say, Lexicon Optimization, to be run along side RIP/CD to address its shortcomings. So we may want to give it the benefit of the doubt when it gets stuck at a grammar whose extension is a superset of the target grammar's extension, a superset language. When a learner fails to even make it to a superset language, however, we can be sure RIP/CD has really failed at a problem it was designed for. A learner who converges on a grammar that yields a superset language of the target will have been successful under the notion of 'superset success'.

We used the criteria of extensional equivalence and superset equivalence in the simulations (sometimes called extensional success and superset success).

5 Simulations

We ran two simulations (Simulations I and II) using constraints related to the voicing in the English plural morpheme (Lombardi 1996), the agreement in nasality between vowels and the coda also in English (Kager 1999:31), and denasalization (of consonants) phenomena found in Mandar, Toba Batak, and Kaingang (Kager 1999:81). The eight constraints considered were:

- *V_{ORAL}N: No oral vowels before a nasal consonant in the coda.
- *$\tilde{V}$: No nasal vowels.
- *N$\underset{\circ}{C}$: No nasal consonants before an unvoiced consonant.
- Ident(voc): Voice faithfulness.
- Ident[V](nas): Nasality faithfulness for vowels.
- Ident[C](nas): Nasality faithfulness for consonants (outranking allows for denasalization of /n/ to /t/).
- *voice: No voiced obstruents.
- *$\underset{\circ}{C}\underset{\smile}{C}$: No unvoiced consonant before a voiced consonant.

Max and Dep were implicitly included as always undominated, meaning that no insertions or deletions of segments were permitted. It can be seen that the constraints interact in unexpected ways. For instance, while *$\underset{\circ}{C}\underset{\smile}{C}$ was proposed for the devoicing of an underlying voiced plural -s following an unvoiced coda (Harms' generalization in Lombardi 1996), it also prevents denasalization of /n/ to /t/ before a voiced consonant.

The simulations were run using the standard RIP/CD procedure and the Variationist RIP/CD procedure, both described earlier. In each section below we report results first for standard RIP/CD and then after a comparison of RIP/CD and Variationist RIP/CD. We

also give an analysis of the Variationist RIP/CD algorithm's failure to learn English at the end of the second simulation.

5.1 Simulation I

5.1.1 Model and Method

The first simulation used only the first six constraints above, those related to nasality agreement and denasalization, plus MAX and DEP which were always considered undominated. The number of possible fully stratified grammars is 6! = 720, though this generates only 16 extensionally unique languages. While only fully stratified grammars were tested as target languages (following Tesar and Smolensky 1996), the learner hypothesizes fully and non-fully stratified grammars along the way. The number of possible stratified grammars is a tad less than 6^6.

The eight underlying forms of all of the target grammars in this simulation had the shape C_1VNC_2. The choice of C_1 did not matter and was constant in all word forms: below we use the consonant /p/. V is either an oral (/æ/) or nasal (/æ̃/) vowel; N is either not present or the nasal consonant /n/; and C_2 is either a voiced (/g/) or unvoiced (/k/) consonant. That generates eight base forms. These forms plus four additional forms where N is realized as a (denasalized) consonant /t/ made up the set of 12 candidates yielded by Gen. In other words, /t/ is an allophone of /n/ that is excluded from the N position in underlying forms.

An English language order of constraints is $*V_{ORAL}N \gg$ IDENT(voc) $\gg$ IDENT[C](nas) $\gg *N\mathring{C} \gg *\tilde{V} \gg$ IDENT[V](nas). This yields the mapping from base forms to the four surface forms shown in Figure 10. This grammar mimics English by excluding the output forms where the nasalization of the vowel and following consonant do not match and the four forms which show denasalization. For comparison, the constraint ordering closest to French ranks the faithfulness constraints higher to preserve contrastiveness in vowel nasality before non-nasal consonants, and preventing denasalization as English does. One such order

is *V_{ORAL}N ≫ Ident(voc) ≫ Ident[C](nas) ≫ *N$\underset{\circ}{\text{C}}$ ≫ Ident[V](nas)≫ *$\tilde{\text{V}}$. We will also report results for a language like English in that vowel nasality is non-contrastive but where denasalization of consonants before unvoiced consonants also occurs. One such language, which we will call Mandar′, is yielded by the order *N$\underset{\circ}{\text{C}}$ ≫ *V_{ORAL}N ≫ Ident(voc) ≫ Ident[C](nas) ≫ *$\tilde{\text{V}}$ ≫ Ident[V](nas).

base	English	French	Mandar′
pæk	pæk	pæk	pæk
pæg	pæg	pæg	pæg
pænk	pæ̃nk	pæ̃nk	pætk
pæng	pæ̃ng	pæ̃ng	pæ̃ng
pæ̃k	pæk	pæ̃k	pæk
pæ̃g	pæg	pæ̃g	pæg
pæ̃nk	pæ̃nk	pæ̃nk	pætk
pæ̃ng	pæ̃ng	pæ̃ng	pæ̃ng

Figure 10: The base forms and outputs of three languages, English, French, and a language based on Mandar which we call Mandar′.

A separate Markov model was created for each of the 16 target languages, and simulations were run for a sufficient number of iterations until the model converged into a nearly stable state.

In addition to different target languages, different sets of initial states were tried — i.e. different notions of UG. The first condition was when the learner started in any of the 6!=720 fully stratified grammars with equal probability. This is one of the most relaxed notions of UG. (One might permit the learner to start with any stratified grammar, though the number of such grammars is practically uncountable so we could not simulate this condition.) We also tested the case where the learner could start in any fully stratified grammar where all of the markedness constraints outranked all of the faithfulness constraints. This conception of UG follows from the work on the subset problem described earlier. There are 36 such grammars in the six-constraint system in this simulation. We denote this conception of UG as the 36 M ≫ F grammars. We also considered a bistratal grammar with all of

the markedness constraints in the top stratum and all of the faithfulness constraints in the bottom stratum, abbreviated {M} ≫ {F}: { *V_{ORAL}N , *$\tilde{V}$, *N$\mathring{C}$ } ≫ { IDENT(voc), IDENT[V](nas) , IDENT[C](nas) }. Finally, we considered a monostratal grammar, in which the constraints are unranked.

5.1.2 Results

The transition probabilities between hypothesis states under a particular target language can be visualized as a state diagram. Because the size of the hypothesis space is on the order of 6! it is impossible to show a complete state diagram, but when the initial hypotheses are restricted to a small subset of the 6! grammars, most hypotheses become unreachable.

We report results first for when the initial hypotheses are restricted to the 36 fully stratified grammars in which all markedness constraints outrank all faithfulness constraints ('M ≫ F'), with English as the target language. This type of initial condition can be found throughout the OT language acquisition literature. The state digram is composed of six disconnected subgraphs of similar structure. One of the subgraphs is depicted in Figure 11 below. The original RIP/CD formulation was used. Each node represents a hypothesis grammar. Solid transition arrows between the nodes indicate the probability of a learner moving from one hypothesis to another after a single word input. Note that they are all directed downward from the initial states in the top row. Dashed transition arrows represent the action of refinement at the end of the RIP/CD algorithm and land on nodes representing the set of refinements that are extensionally equivalent ('e') or superset-equivalent ('es') to the target grammar. (The 'f' and 'int' set of refinement grammars represent those grammars with which the learner establishes the right input–output mapping (f) or interprets all overt forms correctly using Robust Interpretive Parsing — Tesar's (1998a) notion of success (int).) The diagram shows that the learner will necessarily finish RIP/CD on grammar 37 and then has a 50% chance of performing refinement to yield an extensionally equivalent grammar to English, but a 100% chance to refining to a superset of English.

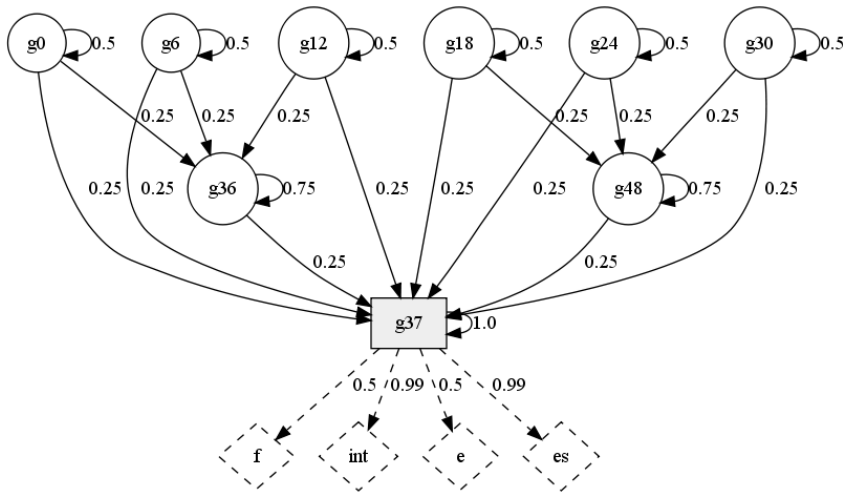

Figure 11: One-sixth of the state transition diagram for the English target language in the first simulation. Each node represents a hypothesis grammar. Solid transition arrows indicate the probability of a learner moving from one hypothesis to another after a single word input. Dashed transition arrows represent the action of refinement at the end of the RIP/CD algorithm and land on nodes representing the set of refinements that are functionally equivalent ('f'), extensionally equivalent ('e'), superset-equivalent ('es'), or interpretively correct ('int').

The state diagram becomes more complex as the target hypothesis requires even-lower-ranked markedness constraints. In a language where nasality is distinctive in all contexts and so nasality faithfulness is ranked high, more steps are required to move the markedness constraints down from their high positions in the initial states, and more hypotheses are on the possible paths from initial grammars to final grammars. The state diagram for such a language is far too large to include as a figure.

Convergence rate for a target language is the probability that a learner halts with a successful hypothesis, with initial hypotheses considered equally probable. Figure 12 shows the rate of convergence for each of the 16 target languages, under the notions of extensional and superset success. Each line is a target language, and each iteration is a step of RIP/CD

for one randomly drawn overt form of the language. Again, just the 36 M $\gg$ F grammars were allowed as initial states. As expected, the rates of convergence are quite rapid, within 20-50 iterations.

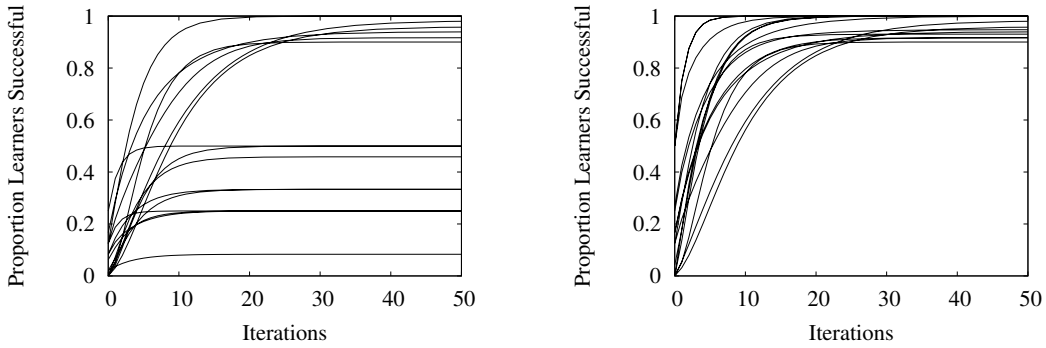

Figure 12: Convergence rates in simulation I for the 16 target languages, under the notion of 'extensional success' (left) and 'superset success' (right). The initial states consisted of the 36 M $\gg$ F grammars.

The number of target languages that had successful convergence with probability at least .9 — i.e. the languages that are learnable — for the different initial state sets and two success criteria considered, is reported in Figure 13. With the extensional equivalence criterion for success, no conception of UG tried allowed all 16 languages to be learned, though the best, at 7 of 16 languages, was the 36 M $\gg$ F conception of UG. In that case the learner always arrived at least at a superset language. For the other conceptions of UG, the learner mostly but not always halted on a superset-equivalent grammar.

A table of the probabilities of convergence to a successful grammar for the English, French, and Mandar′ target languages, again under the different initial state sets and success criteria and after a sufficient number of iterations for the learner to have (virtually) settled on a final hypothesis, is reported in Figure 14. Since these languages are attested (or in the case of Mandar′ likely attested), we can determine which initial state sets and criteria

Initial States	Extensional	Superset
All 720	4	12
36 M $\gg$ F	7	16
{M} $\gg$ {F}	5	12
Monostratal	4	12

Figure 13: Out of the 16 target languages, the number that had successful convergence with probability at least .9, for the different initial state sets and success criteria considered, in simulation I.

are plausible. From this table it can be seen that the monostratal grammar is not a good

Initial States	Success Type	English	French	Mandar′
All 720	Extensional	.19	.33	.07
All 720	Superset	.99	1.00	.93
36 M $\gg$ F	Extensional	.33	.50	.08
36 M $\gg$ F	Superset	1.00	1.00	.92
{M} $\gg$ {F}	Extensional	.00	.50	.00
{M} $\gg$ {F}	Superset	1.00	1.00	.67
Monostratal	Extensional	.08	.33	.00
Monostratal	Superset	.71	1.00	.24

Figure 14: Probability of convergence for learners of three languages in simulation I.

candidate for UG, as even under the most generous success criteria, superset equivalence, English is learned with only a probability of .71. To the extent Mandar′ is a believable human language, it suggests that neither the monostratal nor the bistratal grammars could be the initial state, since the language is learned at best with only a probability of .67. At worst, the probability of learning some of these languages goes below the baseline success rate of $\frac{1}{16} = .0625$, if the learner guesses one of the 16 target languages at random. The markedness-over-faithfulness conception of UG surprisingly does not show any advantage over the least restrictive conception of UG.

We did find that by reducing the set of 36 M $\gg$ F grammars to the 18 grammars that have IDENT[C](nas) ranked below IDENT[V](nas), all target languages were learnable under

the superset notion of success with probability $\geq .95$. From none of these initial states could a learner acquire all of the target languages under the stricter extensional success.

When using the alternate Variationist RIP/CD algorithm described earlier, with the least restrictive notion of UG (all 720 grammars as initial states), learners were no better able to learn languages' extensions perfectly. The number of languages out of the 16 target languages that are learned with probability at least .9 are reported in Figure 15. On the other hand, learners would always learn a superset of every language eventually — all learner's probabilities of reaching a superset-equivalent grammar converged on 1.0 as the number of iterations were carried forward.

	Extensional	Superset
Standard RIP/CD	4	12
Variationist RIP/CD	4	16

Figure 15: Out of the 16 target languages, the number that had successful convergence with probability at least .9, for the different success criteria considered and the two algorithms run, in simulation I. The UG was unrestricted: all 720 fully stratified grammar were permitted as initial states. The first row can be found in Figure 13 as well.

The results of Simulation I were promising for RIP/CD. The unrestricted UG, when any fully stratified grammar could be an initial state, was as good as any other for learning the three known or assumed to be attested languages English, French, and Mandar′. For all three languages the learner arrives at a superset language with high probability. Variationist RIP/CD seemed to fix any language's problems with learning a superset: with this algorithm, all 16 languages could be learned under this generous notion of success. On the other hand, neither the standard nor the Variationist RIP/CD were particularly successful at learning languages' extensions: the highest probability was .5 for French. Variationist EDCD did not make any previously unlearnable language learnable under extensional success.

5.2 Simulation II

5.2.1 Model and Method

The eight constraints considered in Simulation II included all of those listed above, the six from Simluation I plus:

- *VOICE: No voiced consonants.
- *C̥C̬: No unvoiced consonant before a voiced consonant.

(and MAX and DEP always undominated). The 8! = 40,320 fully stratified hierarchies yielded 54 functionally distinct grammars to test as target languages. The languages were generated with the 12 base forms above plus 12 more with the plural suffix, which following Lombardi (1996) is always voiced underlyingly. The output candidates include these and an additional 12 with an unvoiced plural suffix.

An English-like grammar is *V_{ORAL}N ≫ *C̥C̬ ≫ IDENT(voc) ≫ *VOICE ≫ IDENT[C](nas) ≫ *NC̥ ≫ *$\tilde{V}$ ≫ IDENT[V](nas). The Mandar′ grammar in this section exhibits the English oral/nasal vowel and plural voicing patterns but also includes denasalization. It has the same constraint hierarchy but with *NC̥ ranked before the other eight constraints. A French′ grammar is also reported which has all faithfulness constraints ranked highest, exhibiting a nasality vowel contrast everywhere, but neither denasalization nor devoicing of the plural marker. Figure 16 shows the input–output mapping for the 12 new base forms with the plural suffix:

In this simulation, there were far too many permutations of the constraints to simulate all of them as possible initial states. We considered three possibilities for UG: 720 M ≫ F grammars (more in this simulation because there are more constraints), the bistratal {M} ≫ {F} grammar, and the monostratal grammar.

base	English	French′	Mandar′
pækz	pæks	pækz	pæks
pægz	pægz	pægz	pægz
pænkz	pæ̃nks	pænkz	pætks
pængz	pæ̃ngz	pængz	pæ̃ngz
pæ̃kz	pæks	pæ̃kz	pæks
pæ̃gz	pægz	pæ̃gz	pægz
pæ̃nkz	pæ̃nks	pæ̃nkz	pætks
pæ̃ngz	pæ̃ngz	pæ̃ngz	pæ̃ngz

Figure 16: The additional base forms and outputs of three languages, English and languages based on French and Mandar, in simulation II.

5.2.2 Results

Figure 17 shows the rate of convergence for each of the 54 target languages, under the notions of extensional and superset success. Just the 720 M $\gg$ F grammars were allowed as initial states. As with the first simulation, convergence was quite fast, within roughly 30 iterations.

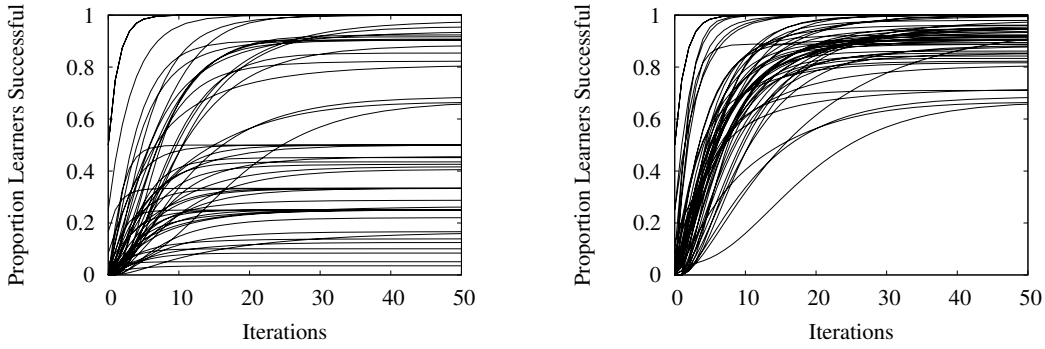

Figure 17: Convergence rates in simulation II for the 54 target languages, under the notion of 'extensional success' (left) and 'superset success' (right). The initial states consisted of the 720 M $\gg$ F grammars.

The number of target languages that had successful convergence with probability at

least .9 — i.e. the languages that are learnable — for the different initial state sets and success criteria considered, is reported in Figure 18. Recall that in the previous simulation the M $\gg$ F grammars as initial states allowed all target languages to be learned under the notion of superset success. With the 720 M $\gg$ F grammars as initial states here, the number of learnable languages under the superset criteria reduced to two-thirds.

Initial States	Extensional	Superset
720 M $\gg$ F	15	36
{M} $\gg$ {F}	15	35
Monostratal	5	13

Figure 18: Out of the 54 target languages, the number that had successful convergence with probability at least .9, for the different initial state sets and success criteria considered, in simulation II.

A table of the probability of convergence for the English, French′, and Mandar′ target languages, again under the different initial state sets and success criteria and after 34 iterations, is reported in Figure 19. Since these languages are attested or expected to be attested, we can determine which initial state sets and criteria are plausible. Recall

Initial States	Success Type	English	French′	Mandar′
720 M $\gg$ F	Extensional	.29	.90	.04
720 M $\gg$ F	Superset	.89	.90	.71
{M} $\gg$ {F}	Extensional	.00	1.00	.00
{M} $\gg$ {F}	Superset	.52	1.00	.50
Monostratal	Extensional	.00	1.00	.00
Monostratal	Superset	.20	1.00	.16

Figure 19: Probability of convergence for learners of three languages in simulation II.

that in the previous simulation, only with all grammars as possible initial states or with the 36 M $\gg$ F grammars as initial states were all three languages learnable under the superset notion of success (learnable again being with probability at least .9). Here English is best learned with a probability of only .89. Note that under the notion of extensional

success, the learner never has a probability more than .29 of learning English or Mandar′, showing that the learner is very often getting trapped by the subset problem. French′ is distinctly more learnable because its faithfulness constraints are ranked highest, and this avoids the problem of demoting faithfulness constraints discussed earlier. It is still not perfectly learnable, however. With the 720 M $\gg$ F grammars as initial states, it is learned with only probability .9 — the same as in the last simulation. At worst, the probability of learning some of these languages goes below the baseline success rate of $\frac{1}{54} = .019$, if the learner guesses one of the 54 target languages at random.

There was no fully stratified hierarchy in the set of 720 M $\gg$ F grammars from which all target languages were learnable under either the extensional or superset success criteria. An exhaustive search of all possible stratified hierarchies was not possible, and as a result unlike in the previous section we could find no UG that supported learning all target languages.

When using the alternate Variationist RIP/CD algorithm, learners were able to learn languages' extensions somewhat better, and as with the previous simulation learners would learn a superset of every language eventually. Here the initial states were the 720 M $\gg$ F constraint rankings. The number of learnable languages out of the 54 target languages are reported in Figure 20, and the convergence rates are graphed in Figure 21.

	Extensional	Superset
Standard RIP/CD	15	36
Variationist RIP/CD	22	54

Figure 20: Out of the 54 target languages, the number that had successful convergence with probability at least .9, for the different success criteria considered and the two algorithms run, in simulation II. The first row can be found in Figure 18 as well.

A table of the probability of convergence for the English, French′, and Mandar′ target languages using Variationist RIP/CD for the same initial state set and extensional success is reported in Figure 22. English and Mandar′ remain unlearnable in this case, with French′ still learnable and now guaranteed to be learned.

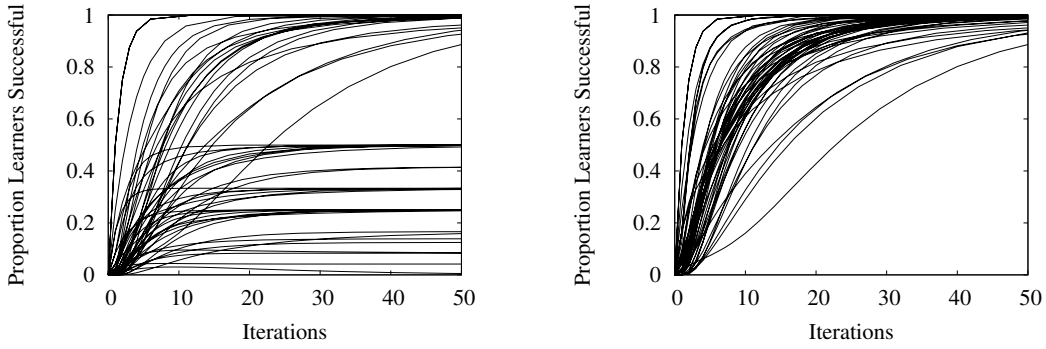

Figure 21: Convergence rates in simulation 2 using Variationist RIP/CD for the 54 target languages, under the notion of 'extensional success' (left) and 'superset success' (right). The initial states consisted of the 720 M $\gg$ F grammars.

	English	French′	Mandar′
Standard RIP/CD	.29	.90	.04
Variationist RIP/CD	.33	1.00	.00

Figure 22: Probability of convergence for learners of three languages in simulation II with standard and Variationist RIP/CD, with the initial states the 720 M $\gg$ F grammars and under extensional success.

The picture of RIP/CD painted by this analysis is not quite so good. Under none of the three conceptions of UG were the three languages in focus learnable with probability at least .9 and the weaker superset criteria for success. A smaller proportion of all of the languages in the factorial typology were learnable as well. As in the past simulation, all languages were learnable with Variationist RIP/CD under the notion of superset success. Variationist RIP/CD also made more languages learnable under the strongly extensional criteria for success, though neither English nor Mandar′ were substantially helped.

5.2.3 Analysis

To see why Variationist RIP/CD still does not completely solve the language learning problem, let us take the case of English with UG being the 720 M ≫ F grammars. We reported above that in this case the learner has only a 33% chance of, in the end, having a grammar that is extensionally equivalent to the target language. The Markov model for this learner contains 1,048 states, far too large to visualize here, so we will use as an example to see where the problem is one particular path from an initial state to when the learner makes no more changes to his hypothesis, and then performs refinement to a grammar that does not extensionally match English.

We will start with the initial state *NC̥ ≫ *$\tilde{\text{V}}$ ≫ *V_{ORAL}N ≫ *C̥C̥ ≫ *VOICE ≫ IDENT(voc) ≫ IDENT[C](nas) ≫ IDENT[V](nas), which has all of the markedness constraints ranked above all of the faithfulness constraints. When the learner hears /pægz/, he constructs the tableau shown partially in Figure 23. (The Robust Interpretive Parsing step is omitted for brevity.) He then demotes *VOICE (the constraint violated by the negative candidate but not the positive candidate) below IDENT(voc) (the constraint violated by the positive candidate but not the negative candidate), which puts *VOICE into the same stratum as IDENT[C](nas). This first step is in the right direction. One fourth of the overt tokens result in this change.

pægz	...	*C̥C̥	*VOICE	IDENT(voc)	IDENT[C](nas)	...
pægz			*!*			
☞pæks				**		

Figure 23: The first step in our example that leads to an eventual failure. The positive candidate is /pægz/, the result of Robust Interpretive Parsing, while the negative candidate is /pæks/, the winner of the tableau according to the learner's current grammar. *VOICE is about to be demoted to the stratum immediately below IDENT(voc).

The second and final step of constraint demotion occurs when the learner hears the

output token /pə̃nk/. Robust Interpretive Parsing infers the base form is, also, /pə̃nk/. Under the learner's current grammar, the optimal output for this base form is /pætk/. This differs from what it heard, /pə̃nk/, and the learner demotes *NC̥ and *Ṽ to the stratum immediately below IDENT[C](nas) (shown partially in Figure 24). This corrects the fact that the denasalization and *Ṽ constraints were both ranked too high: both must be dominated by consonant nasality faithfulness. Again, one fourth of the overt tokens result in this change.

pə̃nk	*NC̥	*Ṽ	*V_{ORAL}N	...	*VOICE	IDENT[C](nas)	IDENT[V](nas)
pə̃nk	*!	*					
☞pætk						*	*

Figure 24: The second step in our example that is the crucial step leading to the eventual failure. The positive candidate is /pə̃nk/, the result of Robust Interpretive Parsing, while the negative candidate is /pætk/, the winner of the tableau according to the learner's current grammar. *NC̥ and *Ṽ are about to be demoted to the stratum immediately below IDENT[C](nas).

The learner now has the stratified hypothesis *V_{ORAL}N ≫ *C̥C̥ ≫ IDENT(voc) ≫ { *VOICE, IDENT[C](nas) } ≫ { *NC̥ , *Ṽ , IDENT[V](nas) }, which (as written) has exactly the same order of constraints as a licit English grammar except for the grouping of constraints into strata. This is the last hypothesis the learner will ever posit — he is stuck. This is one of just 12 states from which the learner cannot escape. (There were no cycles, i.e. the learner indefinitely alternating between two or more hypotheses, in any of the simulations.) We will see below that the crucial ordering *Ṽ ≫ IDENT[V](nas) in the English grammar cannot be achieved because it would require demoting the IDENTITY constraint out of the final stratum. We have already established that IDENTITY constraints are never demoted in an OT system like this one: the learner never posits an IDENTITY violation in the positive candidate.

From any of the 12 final trap states the learner may end up in, all of which are not fully

stratified, the learner performs refinement to pick a fully stratified grammar. These states vary in the probability that the refinement chosen is extensionally equivalent to the target: some yield no refinement that matches English, some have a probability of one half, and some always yield a refinement that matches English. In the path we have been following, the final hypothesis has a 50/50 chance of yielding a successful refinement.

One refinement of this final grammar has the last two constraints reversed as compared to English. For the base form /pæ̃g/, this order yields /pæ̃g/ as the optimal candidate rather than the English form /pæg/. This is demonstrated in the tableau in Figure 25.

pæ̃g	*V_{ORAL}N	*C̥C̆	...	IDENT[V](nas)	*Ṽ
pæg				*!	
☞pæ̃g					*

Figure 25: A refinement of a sink state generates an incorrect English form, but yields a superset of English.

This example starts off with the best of circumstances for the learner: the faithfulness constraints are ranked low, he is in the environment of a language we know to be attested, and we equip him with the corrected Variationist RIP/CD algorithm. However, the learner gets stuck prematurely. He fails to revise his grammar to make a critical ordering because he cannot get the data he needs to demote a faithfulness constraint. In the end, he is left with a stratum that can be refined one way to match the target language and another way to yield a superset of the target language. The learner fails to learn English.

6 Conclusion

We began this thesis by examining RIP/CD at an abstract level. As noted in the introduction, it is the step of Robust Interpretive Parsing, the step of inferring hidden structure, rather than the Constraint Demotion component that is of interest. EDCD and Variationist EDCD have shown that Constraint Demotion itself is sufficient for solving the learning problem if hidden structure is provided to the learner. When hidden structure is taken away, how well does Robust Interpretive Parsing fill in the gap?

Most of what had been known about RIP/CD had been based on OT analyses of metrification. These OT systems present a special case for RIP/CD, however. Outputs, what the learner hears from his environment, tell the learner immediately what the base form of the token was. The learner must infer a structural description (location of feet), but not the base form. We turned in this thesis to OT systems for segmental alternations. These systems are a special case too, but from the opposite direction. Here we cut out phonetics so that the learner hears the surface phonological form directly. The learner need not infer the surface phonological form then, but instead must infer the base form of tokens he hears, which could be substantially different from the surface phonological form.

These two types of OT analyses are quite different in terms of their ramifications for RIP/CD. Some of RIP/CD's limitations that have been raised before — when Robust Interpretive Parsing chooses structures that cannot be optimal and cycles induced by this case — do not arise when structures do not need to be inferred. However, new problems can arise and other problems remain. We discussed these problems formally and evaluated their impact in two simulations involving constraints related to oral and nasal vowel alternations, the English plural morpheme's voicing, and denasalization of consonants.

First, the problem of the final refinement step in EDCD raised by Boersma (2008b) persisted in RIP/CD. Our simulations showed that switching from the original RIP/CD to a Variationist RIP/CD based on Boersma's suggestions for EDCD gave the learner a better chance of learning languages. In the second simulation, the learner was able to

learn more languages precisely, that is finding an extensionally equivalent grammar, and in both simulations the Variationist RIP/CD learner was guaranteed to converge on a superset-equivalent grammar for all target languages. If we put the subset problem aside, Variationist RIP/CD indeed seems to have solved the OT language learning problem.

But the subset problem is another ongoing problem for RIP/CD. It is, however, not merely a matter of lacking negative evidence. (Variationist) EDCD is not given negative evidence either but it is claimed to have solved the (narrower) language learning problem. RIP/CD generates its own negative evidence, but its negative evidence is imperfect. We showed that Robust Interpretive Parsing can generate faulty negative evidence, at least when there are unusual I–O constraints like C_1 on page 21, and leave open for further research whether constraints such as MAX, DEP and LINEARITY are 'unusual' enough to break Robust Interpretive Parsing. In our simulations, this did not occur. On the other hand, Robust Interpretive Parsing can fail to generate enough negative evidence. In particular, it fails to generate the kind of evidence needed to demote IDENTITY constraints.

The question of how to rank identity constraints in UG so as to avoid this problem had been taken up before — i.e. constrain UG so they start off ranked low. In Simulation I, this type of ranking improved the rate learners would learn languages, but did not solve the problem. Neither English, nor French, nor the hypothesized Mandar-based language were learned (that is, arriving at an extensionally equivalent grammar) with any reasonable probability. Further, no markedness-over-faithfulness ranking, in either simulation, allowed all target languages to be learnable. Constraining UG does not appear to render the subset problem solved.

Comparing the results of the two simulations we see what may be the beginning of a trend. As the number of constraints grows, the probability that a learner successfully learns his target language decreases. The probability of extensional success for an English learner is lower for each conception of UG in the second simulation compared to the first, and likewise for the number of learnable languages out of the factorial typology. This is

probably from the fact that with more I–O constraints to rank, fewer initial states put the learner on the right track by coincidentally having those constraints ranked correctly from the beginning. Far more faithfulness constraints are proposed in the literature than have been considered here, raising the question of whether for any given target language the learner can start off anywhere but finished.

This thesis provides a foundation for moving forward with learnability research in Optimality Theory, especially as it relates to the subset problem, segmental alternations, and I–O constraints. We now have baseline numbers to compare future models against, to judge whether they face the same set of problems, whether they make any headway against the subset problem, or whether they make different predictions about which members of the factorial typology of languages are learnable.

Appendix

A Details of the Markov Model and Simulations

The following notes explain the details of the Markov model used to simulate RIP/CD learners. A $n \times n$ transition matrix M is constructed such that M_{ij} is the transition probability from state i to state j, that is, the proportion of learners with current hypothesis grammar i changing their hypothesis to grammar j after processing one token from the target language. M_{ij} is computed by running one iteration of RIP/CD for each overt form of the target language on a learner in state i and determining the proportion of times RIP/CD takes the learner to state j, weighted by the probability of encountering the overt forms. As we noted in section 2.4, RIP/CD can be nondeterministic when the hypothesis grammar is not fully stratified and multiple interpretive parses and optimal candidates arise. In that case, we go through each choice of parse and optimal candidate in turn to complete the iteration of RIP/CD, treating each parse–candidate pair as equally probable.

Formally:

$$M_{ij} = \sum_{w \in \mathcal{L}} p^{\mathcal{L}}(w) \cdot \text{RIP/CD}(i, w, j)$$

where w is a overt form drawn from the target language $\mathcal{L}$, $p^{\mathcal{L}}(w)$ is its probability of being encountered by the learner, and $\text{RIP/CD}(i, w, j)$ is the probability that one iteration of RIP/CD starting with hypothesis i and using token w takes the learner to hypothesis grammar j. When the hypothesis grammar i is fully stratified the RIP/CD algorithm is deterministic and $\text{RIP/CD}(i, w, j)$ is either 0 or 1. This was computed from a richness-of-the-base-style assumption. Note that, for any given i and, in the first line, w,

$$\sum_j \text{RIP/CD}(i, w, j) = 1,$$

$$\sum_{w \in \mathcal{L}} p^{\mathcal{L}}(w) = 1, \text{ and}$$

$$\sum_j M_{ij} = 1$$

If l_k is a row vector of length n giving a learner's probability of being in each hypothesis state at iteration k, then $l_{k+1} = l_k \cdot M$. Similarly, if D_k is a $n \times n$ matrix where $(D_k)_{ij}$ gives the probability that a learner who started at iteration zero in state i is k iterations later in state j, then $D_{k+1} = D_k \cdot M$. $D_0 = I$, the identity matrix, since a learner who started in state i is in state i at iteration zero, and so $D_k = I \cdot M^k = M^k$, yielding a complete description of the distribution of learners after k tokens.

The hypotheses considered when building the transition matrix could be all possible stratified grammars. When there are more than a handful of constraints the number of stratified grammars grows unreasonably large. In practice, the transition matrix is created dynamically to include only those hypothesis states that are reachable from some small set of chosen initial states. As a result, separate simulations are run for difference conceptions of UG.

The RIP/CD algorithm is detailed earlier and elsewhere. In short: the learner hears an output o from the environment. He uses Robust Interpretive Parsing to infer the base form b and candidate c corresponding to o, according to his current hypothesis grammar g. If Robust Interpretive Parsing yields more than one option, the learner chooses one at random. The learner than runs b through a tableau under g and computes the optimal candidate c'. If multiple candidates are optimal, the learner chooses one at random. (Tesar and Smolensky (1996, 2000) did not say which of the multiple optimal candidates should be used. In Tesar (1998a), the learner picks one in a way determined arbitrarily by the simulation implementation, which was not described in enough detail to be replicable. They did not note that a similar problem can occur during Robust Interpretive Parsing. Boersma 2003:440 noticed this as well and adopted the same change.) In Variationist RIP/CD, a randomly selected refinement of g (described below) is used for the previous Robust Interpretive Parsing and tableau solving steps. If c and c' differ, the learner updates g by moving all constraints violated by c but not c' to the stratum immediately below the highest stratum containing a constraint violated by c' but not c. In non-fully stratified hierarchies,

tableaux are evaluated by pooling the violations of all of the constraints in a stratum (as in Tesar and Smolensky 1996, 2000 but not Tesar 1998b).

The last step in RIP/CD, as noted earlier, is for the learner to refine his possibly non-fully stratified grammar so that it is fully stratified, as the learner is assumed to know that all adult grammars are fully stratified. A refinement is computed by replacing each (non-singleton) stratum s with individual strata for each of the constraints in s, in any order. For example, the grammar {A,B} $\gg$ {C,D} has four refinements: A $\gg$ B $\gg$ C $\gg$ D, B $\gg$ A $\gg$ C $\gg$ D, A $\gg$ B $\gg$ D $\gg$ C, B $\gg$ A $\gg$ D $\gg$ C. Since the choice of refined grammar matters, we assume the learner chooses a refined grammar at random. Unfortunately, we cannot always include all possible choices in the simulation because as the number of constraints increases, the number of refinements grows super-exponentially. Instead, we draw a relatively small random sample of refinements for each final hypothesis grammar the learner chooses. In Variationist RIP/CD simulations, we used this same procedure prior to Robust Interpretive Parsing.

Each refinement is compared to the target grammar on two notions of success — extensional equivalence and and superset equivalence. (Here again a 'richness' assumption is made about the lexicon: in comparing the learner's hypothesis grammar with the target grammar, it is assumed that the learner's lexicon is the same as the adult lexicon, namely the set of all possible base forms.) We will note as q_j^S the probability that a learner who has hypothesis j chooses a refinement that is S-equivalent to the target grammar (which we compute by merely iterating over the computed refinements of each grammar).

The probability that a learner who started in state i chooses a refinement S-equivalent to the target grammar after k iterations is then

$$p^S(i,k) = \sum_j (M^k)_{ij} \cdot q_j^S$$

and if we consider all grammars in UG to be equally probable initial states, then the

probability of successfully learning a grammar is

$$p^S(k) = \frac{1}{|UG|} \sum_{i \in UG} p^S(i,k)$$

We can carry out k indefinitely until it appears that the value of $p^S(i,k)$ has stopped changing, i.e. when a sufficient number of iterations have been run so that learners appear to have converged. One can then ask how many initial states had learners that converged to a successful hypothesis with high probability, or if the initial states are considered equally probable, then what the probability $p^S(k)$ is that any learner converges to a successful grammar. A probability that isn't high enough suggests either that the target language is unlearnable from a particular initial state or any initial state, in the latter case making the empirical prediction that the language does not occur, or in the former case that the initial state is not available in UG.

This entire process is then repeated for each target language the learner may be exposed to.

References

Alderete, John. 2008. Using learnability as a filter on computable functions: A new approach to Anderson and Browne's generalization. *Lingua* 118:1177–1220.

Alderete, John, and Bruce Tesar. 2002. Learning covert phonological interaction: An analysis of the problem posed by the interaction of stress and epenthesis. RuCCS-TR-72, Rutgers Center for Cognitive Science.

Apoussidou, Diana, and Paul Boersma. 2004. Comparing two Optimality-Theoretic learning algorithms for Latin stress. In *Proceedings of WCCFL 23*, ed. B. Schmeiser, V. Chand, A. Kelleher, and A. Rodriguez, 101–114. Somerville, Mass.: Cascadilla Press.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21:43–58.

Boersma, Paul. 2003. "Bruce Tesar and Paul Smolensky (2000). Learnability in Optimality Theory.". *Phonology* 20:436–446.

Boersma, Paul. 2008a. Emergent ranking of faithfulness explains markedness and licensing by cue. ROA 954-0308.

Boersma, Paul. 2008b. Some correct error-driven versions of the constraint demotion algorithm. ROA 980-0708.

Boersma, Paul, and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45–86.

Boersma, Paul, and David Weenink. 2008. Praat: Doing phonetics by computer [computer program].

Demuth, Katherine. 1995. Markedness and the development of prosodic structure. In *Proceedings of NELS 25*, ed. J. Beckman. GLSA, University of Massachusetts, Amherst.

Fodor, Janet Dean. 1998. Unambiguous triggers. *Linguistic Inquiry* 29:1–36.

Fodor, Janet Dean, and W. G. Sakas. 2005. The subset principle in syntax: Costs of compliance. *Journal of Linguistics* 41:513–569.

Gibson, Edward, and Ken Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.

Gnanadesikan, Amalia. 1995. Markedness and faithfulness constraints in child phonology. Ms., University of Massachusetts, Amherst.

Gold, E. M. 1967. Language identification in the limit. *Information and Control* 10:447–474.

Hale, Mark, and Charles Reiss. 2003. The subset principle in phonology: why the tabula can't be rasa. *Journal of Linguistics* 39.

Hayes, Bruce. 2001. Phonological acqusition in Optimality Theory: The early stages. In *Constraints in phonological acquisition*, ed. Rene Kager, Joe Pater, and Wim Zonneveld, chapter 5, 158–203. Cambridge University Press.

Kager, Rene. 1999. *Optimality Theory*. Cambridge University Press.

Levelt, Clara C., Niels O. Schiller, and Willem J. Levelt. 2000. The acquisition of syllable types. *Language Acquisition* 8:237–264.

Lombardi, Linda. 1996. Restrictions on direction of voicing assimilation: An OT account. *University of Maryland Working Papers in Linguistics* 6.

McCarthy, John. 2005. Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* 4:19–55.

Niyogi, Partha, and Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61:161–193.

Pater, Joe. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39:334–345.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report RuCCS-TR-2, Rutgers Center for Cognitive Science; Also Prince and Smolensky (2000): ROA 537-0802.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, ed. R. Kager, J. Pater, and W. Zonneveld. Cambridge, UK: CUP.

Riggle, Jason. to appear. The complexity of ranking hypotheses in Optimality Theory. *Computational Linguistics* .

Smith, Jennifer L. 2000. Positional faithfulness and learnability in Optimality Theory. In *Proceedings of ESCOL 99*, ed. Rebecca Daly and Anastasia Riehl, 203–214. CLC Publications.

Smolensky, Paul. 1996. The initial state and 'richness of the base' in Optimality Theory. Technical report JHU-CogSci-96-4, Johns Hopkins University Department of Cognitive Science.

Stabler, Edward P. 2008. Computational models of language universals: Expressiveness, learnability and consequences. In *Language universals*, ed. M. Christiansen, C. Collins, and S. Edelman. Oxford University Press.

Tesar, Bruce. 1997. Multi-recursive constraint demotion. ROA 197-0597.

Tesar, Bruce. 1998a. An iterative strategy for language learning. *Lingua* 104:131–145.

Tesar, Bruce. 1998b. Robust interpretive parsing in metrical stress theory. In *Proceedings of the 17th West Coast Conference on Formal Linguistics*, ed. K. Shahin, S. Blake, and E.-S. Kim, 625–639. CSLI.

Tesar, Bruce, and Paul Smolensky. 1996. Learnability in Optimality Theory (long version). Technical Report ROA 156-1196, Rutgers University/Johns Hopkins University.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. MIT Press.

Wexler, Kenneth. 1978. A formal theory of language acquisition. In *ACM 78: Proceedings of the 1978 annual conference*, 409–413. New York, NY, USA: ACM.