University of Pennsylvania Working Papers in Linguistics

Volume 15
Issue 1 *Proceedings of the 32nd Annual Penn Linguistics Colloquium*

Article 25

3-23-2009

Goldilocks Meets the Subset Problem: Evaluating Error Driven Constraint Demotion (RIP/CD) for OT language acquisition

Joshua Tauberer
University of Pennsylvania

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/pwpl/vol15/iss1/25
For more information, please contact libraryrepository@pobox.upenn.edu.

Goldilocks Meets the Subset Problem: Evaluating Error Driven Constraint Demotion (RIP/CD) for OT language acquisition

This working paper is available in University of Pennsylvania Working Papers in Linguistics: http://repository.upenn.edu/pwpl/vol15/iss1/25

Goldilocks Meets the Subset Problem: Evaluating Error Driven Constraint Demotion (RIP/CD) for OT language acquisition

Joshua Tauberer*

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). 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 paper. In the RIP/CD algorithm, the learner starts off with an initial grammar, a somewhat random hypothesis for what the adult grammar is. The learner then reranks constraints in his grammar over time, by demoting them, in response to tokens he hears in his environment until his grammar is compatible with all of the tokens he is exposed to.

The point we hope to make in this paper is that when evaluating RIP/CD, we, like Goldilocks from the children's story, are at pains to find a basis of evaluation that is neither too strong nor too forgiving, but is "just right."

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. RIP/CD succeeds on the first evaluation: it is indeed psychologically plausible.

However, examples exist where RIP/CD fails to make a guess at all or fails to arrive at the right grammar: (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. It has been 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. The success of RIP/CD for OT systems of metrical phonology have been considered on several occasions (Tesar, 1998a; Tesar and Smolensky, 2000; Apoussidou and Boersma, 2004), and in systems involving faithfulness violations but in learning algorithms besides 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 RIP/CD (Smith, 2000; Boersma, 2003).

We ask in this thesis "how bad is it?", specifically for learning plausible natural languages involving input–output faithfulness violations. It is difficult to answer this question, however. There are simply too many degrees of freedom. If a language is not learnable, we say that explains why it is not attested (Boersma, 2003; Alderete, 2008). If it *is* attested, we say that we must have given the learner the wrong initial hypothesis. If we can't find an initial hypothesis that works, we may blame it on the subset problem, a general problem in language learning that RIP/CD wasn't specifically designed to address. 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.

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

*I thank Charles Yang for his direction and honest perspective of the field, Gene Buckley for feedback, Catherine Lai, the whole Phonetics Lab's Splunch group, and the audience of PLC 32 for their feedback. My master's thesis is an extended version of this paper.

and Berwick (1996), rather than using simulations that randomly select a particular order of presentation of stimuli for the learner.

- 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.
- looks across the complete factorial typology of languages predicted by an OT system involving faithfulness violations and sees how many of those can be learned by a RIP/CD learner.

2 Review of RIP/CD

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, the learner may change his hypothesis, but never necessarily knowing whether he has ever finally arrived at the *right* hypothesis.

Tesar and Smolensky improved on the essentially random movements between hypotheses made in the Triggering Learning Algorithm (TLA; Gibson and Wexler, 1994; Niyogi and Berwick, 1996) from the syntactic acquisition literature by taking advantage of the fact that in OT one can see which constraints are at play in the analysis of a linguistic object. 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 1) using the learner's current hypothesis grammar to parse input *and* infer the input's hidden structure, 2) generating (possibly faulty) implicit negative evidence, and 3) updating the grammar to rule out this negative evidence and rule in the positive evidence, using their hidden structure.

We now very briefly review the RIP/CD procedure by way of an example. It is described in detail elsewhere (Tesar and Smolensky, 1996, 2000; Tesar, 1998b). 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. Let us place a learner in the environment of English, with adults ranking the relevant constraints as $*\mathrm{V}_{\mathrm{ORAL}}\mathrm{N} \gg *\tilde{\mathrm{V}} \gg$ IDENT[V](nas) (described in section 4.1). 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 $*\tilde{\mathrm{V}} \gg$ IDENT[V](nas) $\gg *\mathrm{V}_{\mathrm{ORAL}}\mathrm{N}$, which admits in outputs only oral vowels, regardless of a following nasal consonant.

The rest of RIP/CD is a repeating process.

2.1 Robust Interpretive Parsing

The learner hears an overt, output token /pæ̃nt/. The learner next guesses the base form and the full structural description or candidate that the adult used to generate the token. This is the process of Robust Interpretive Parsing. So how can a learner know the underlying form of a phoneme that undergoes alternation without already knowing the cause of the alternation? Of course, if the learner could do this step accurately he would have already pretty much solved the problem of language acquisition.

In RIP/CD, 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 and a grammar the base form and candidate that underlie the output (Tesar and Smolensky, 1996:10). The base form and candidate pair yielded by Robust Interpretive

Parsing satisfies two conditions: 1) the candidate is compatible with the output the learner actually heard, and 2) compared to all other such pairs of base forms and candidates, it is the most harmonic under the grammar.

Our learner, who has just heard /pæ̃nt/, now performs Robust Interpretive Parsing. He assembles a special tableau like the one in Figure 1, according to his current hypothesis grammar. 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 of the pairs are all the same: the output the learner heard. The learner sees that /pæ̃nt/ is the optimal base form under his grammar, which happens to be a correct parse of English /pæ̃nt/.

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

Figure 1: Robust Interpretive Parsing of the output /pæ̃nt/.

2.2 Constraint Demotion

The learner next runs the base form that he computed through Robust Interpretive Parsing through a tableau according to his current hypothesis grammar. This will yield an optimal candidate *for him*. See Figure 2.

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

Figure 2: The learner runs the guessed base form through a tableau according to his current hypothesis grammar and concludes that /pænt/ is optimal (for him).

The learner concludes that /pænt/ is optimal (for him). Now, the learner believes, according to Robust Interpretive Parsing, that in the adult grammar the optimal candidate was actually /pæ̃nt/, since that is what he heard. Let's call it the positive candidate (Tesar's "winner"). This is not the same as what his grammar yields, which we'll call the negative candidate (Tesar's "loser"), so he will update his grammar using constraint demotion to make the positive candidate optimal.

CD is a process of reordering constraints. It says to demote constraints that the positive but not the negative candidate violated so that they are all 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 example, the positive candidate /pæ̃nt/ uniquely violated $*\tilde{V}$, while the negative candidate /pænt/ uniquely violated IDENT[V](nas) and $*V_{\text{ORAL}}N$. CD then says to move the first constraint into the stratum immediately below IDENT[V](nas). That yields a new grammar shown in the tableau below, which also shows that the positive example is now optimal. Note that $*\tilde{V}$ and $*V_{\text{ORAL}}N$ are now not mutually ranked: they are in the same stratum.

No action is taken by the learner when the negative and positive candidates are the same, or if there are no constraints uniquely violated by the positive and negative candidates.

The Robust Interpretive Parsing and constraint demotion procedures repeat with new tokens from the environment until the learner no longer makes any changes to his hypothesis grammar.

Once a grammar has been settled on, the final step of RIP/CD is "refinement" (Tesar and Smolensky, 2000:49). If the learner has ended up with a stratified hierarchy, as in this example, the learner must turn his hypothesis grammar into a fully-stratified hierarchy. Refinement means

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

Figure 3: After constraint demotion, the learner has this grammar, in which the positive candidate comes out optimal.

choosing any totally ranked grammar that preserves all of the ordering relations. For instance, the grammar $\{A,B\} \gg \{C,D\}$ has four refinements: $A \gg B \gg C \gg D$, $B \gg A \gg C \gg D$, $A \gg B \gg D \gg C$, $B \gg A \gg D \gg C$. (See Boersma, 2008 for the ramifications for this step, which is discussed as well in my master's thesis.) Because the choice of refinement matters, we assume the learner chooses one at random.

2.3 Other Algorithms for OT Grammar Learning

Several newer algorithms have been proposed in the literature for OT language acquisition, but we feel it is premature to end the research program on RIP/CD. 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 memoryless nature of RIP/CD, and is clearly psychologically implausible at face value. Multi-Recursive Constraint Demotion (MRCD; Tesar, 1997) carries out a batched process over multiple hypotheses simultaneously but also has serious questions of psychological plausibility, and was not intended to be used when the underlying forms of tokens are not transparent.

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. In repositioning constraints, adjustments to rank are made in small changes on the continuous scale. Boersma (2003) and Apoussidou and Boersma (2004) claim improvements over a comparable CD-based algorithm. GLA, however, assumes the learner knows the underlying forms of what he hears, and we expect that the same problems described in this paper for RIP/CD would apply to "RIP/GLA" (Boersma, 2003) as well.

3 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, evidence of forms *in* the language but not evidence that some forms are *not* in the language, as with RIP/CD, 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 problem in OT was recognized at the origins of constraint demotion, but neither Tesar and Smolensky (1996) nor another head-on attack of the subset problem in Smolensky (1996) adequately explained why it is a problem for RIP/CD. They explain that if the learner posits highly-ranked faithfulness constraints, he may be hypothesizing a superset language. Highly ranked faithfulness preserves contrasts in the base that would otherwise be neutralized to forms that instead satisfy markedness constraints, and so high-faithfulness introduces new forms relative to a grammar with lower-ranked faithfulness but does not necessarily exclude any forms. (Consider the interplay between IDENT(nas) and *Ṽ. Ranking the first over the second preserves nasality contrasts in the base, but the second over the first enforces a markedness constraint and yields a subset of the forms allowed by the first.) Attributing the following to Alan Prince, Tesar and Smolensky suggest that initial hypotheses could be required to rank faithfulness constraints lowest. Thus in the absence of evidence *for* contrasts that require high-ranked faithfulness, the learner would start in and be left with the most restrictive grammar.

The subset problem is a problem for any algorithm that doesn't make use of negative evidence to winnow to the hypothesized language, but as we've seen, CD algorithms work by generating negative evidence. One has to ask, then, why RIP/CD's generated negative evidence isn't good enough.

The answer is that Robust Interpretive Parsing will never suppose a faithfulness violation when it can avoid it. If the learner hears /pæ̃nt/ and both /pæ̃nt/ and /pænt/ could be its base form, Robust Interpretive Parsing will choose the former since it has fewer (in this case IDENTITY) violations when paired with the output /pæ̃nt/. This is true *even if the adult had /pænt/ in mind as the lexical entry*, for which the output incurs an IDENTITY violation. It is only because of Robust Interpretive Parsing, and not a fault of CD generally, that the learner doesn't see all of the relevant violations in the language. If he never encounters a violation of a constraint, he will never get the opportunity to demote it: the learner thus can never demote a faithfulness constraint. If a faithfulness constraint starts ranked high, the learner has no hope to learn a language where it is lower ranked.

The research program among most of the authors that have considered problems for CD 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 grammar—perhaps there are particular initial grammars 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, these constraints must start off in the initial grammar ranked below all markedness constraints, and Smith (2000) suggested further that positional faithfulness constraints must outrank other faithfulness constraints in the initial grammar. 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? (I also suggest in my master's thesis that 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.)

Given the general nature of the subset problem, the fact that Tesar and Smolensky did not design RIP/CD specifically to address it, and that it may be that we just haven't yet figured out the UG that solves the problem, we may want to give RIP/CD some leeway. We return to this in section 4.3.

4 Simulations

4.1 The Linguistic Analysis

We ran several simulations 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$_{\text{ORAL}}$N	No oral vowels before a nasal consonant in the coda.
*Ṽ	No nasal vowels.
*NC̥	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 consonants.
*C̥C̬	No unvoiced consonant before a voiced consonant.

MAX and DEP were implicitly included as always undominated. It can be seen that the constraints interact in unexpected ways. For instance, while *C̥C̬ 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 sixteen underlying forms of all of the target grammars in this simulation had the shape C_1VNC_2(z). 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. A final voiced consonant /z/ for the plural morpheme is optional. That generates sixteen base forms. These forms plus additional forms where N is realized as a (denasalized) consonant /t/ made up the larger set of candidates yielded by Gen.

An English-like grammar is $*\mathrm{V}_{\mathrm{ORAL}}\mathrm{N}$ $\gg$ *C̥C̬ $\gg$ IDENT(voc) $\gg$ *voice $\gg$ IDENT[C](nas) $\gg$ *NC̥ $\gg$ *Ṽ $\gg$ IDENT[V](nas). This yields the mapping from base forms to the surface forms shown in Figure 4. This grammar mimics English by excluding the output forms where the nasalization of the vowel and following consonant do not match and the forms which show denasalization. For comparison, a constraint ordering close to French (French′) maximally ranks the faithfulness constraints to preserve contrastiveness in vowel nasality in outputs, and preventing both devoicing of the plural morpheme and denasalization. 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 a constraint order the same as English but with *NC̥ ranked before the other eight constraints.

base	English	French′	Mandar′	base	English	French′	Mandar′
pæk	pæk	pæk	pæk	pækz	pæks	pækz	pæks
pæg	pæg	pæg	pæg	pægz	pægz	pægz	pægz
pænk	pæ̃nk	pænk	pætk	pænkz	pæ̃nks	pænkz	pætks
pæng	pæ̃ng	pæng	pæ̃ng	pængz	pæ̃ngz	pængz	pæ̃ngz
pæ̃k	pæk	pæ̃k	pæk	pæ̃kz	pæks	pæ̃kz	pæks
pæ̃g	pæg	pæ̃g	pæg	pæ̃gz	pægz	pæ̃gz	pægz
pæ̃nk	pæ̃nk	pæ̃nk	pætk	pæ̃nkz	pæ̃nks	pæ̃nkz	pætks
pæ̃ng	pæ̃ng	pæ̃ng	pæ̃ng	pæ̃ngz	pæ̃ngz	pæ̃ngz	pæ̃ngz

Figure 4: The base forms and outputs of three languages, English and languages based on French and Mandar.

4.2 Simulating RIP/CD with a Markov Model

A Markov model captures the complete behavior of a learner, computing all possible paths the learner may take in response to all possible orders of surface forms, and improves on the simulation methods of Tesar and Smolensky (2000) and others which choose a few particular initial states and randomly generate just a few particular orders of stimuli to feed the learner. The results of a Markov model, i.e., the computed probabilities of successful learning, are exact, whereas with simulations the effects of the random choices made by the simulation are unclear.

A Markov model is described by a state transition diagram. Here, each of the states 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 represent an action after hearing a token from the environment, and the probability of each transition 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 run using a new Python program written by the author.

We considered as target languages all of the languages that arise from all 8! = 40,320 permutations of the constraints; here there were 54 unique languages. 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 is computed by running each base form through a tableau one by one and noting the resulting output.

In addition to different target languages, different sets of initial states were tried—i.e., different notions of UG. There were far too many constraint orders to simulate all of them. The first condition was when 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 720 such grammars in the eight-constraint system in this simulation. We denote this conception of UG as the 720 M $\gg$ 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 $\{M\} \gg \{F\}$: { *V$_{\mathrm{ORAL}}$N , *Ṽ , *NC̥, *voice, *C̥C̬ } $\gg$ { IDENT(voc), IDENT[V](nas) , IDENT[C](nas) }. Finally, we considered a monostratal grammar, in which the constraints are unranked.

Further details of the simulations are presented in my master's thesis.

4.3 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 a correct hypothesis, but it may also be a hypothesis that is clearly wrong but which RIP/CD cannot get the learner out of. There are several situations when RIP/CD can get prematurely stuck, for instance when a faithfulness constraint is ranked too high, as discussed above.

When has a learner that's gotten stuck 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, and say a learner is successful if he converges on that very same order of constraints. This criterion is unrealistically narrow since it is common for different orders of constraints to be essentially equivalent, and so indistinguishable to the learner.

A less strict notion of success is if the learner arrives at a grammar whose yield (i.e., outputs) is the same as the target's yield, i.e., if the grammars have the same "extensions." An extensionally successful learner will appear to speak the same language, though the learner's surface forms may correspond to the wrong lexical entries.

However, even extensional success may not be fair to RIP/CD. We know RIP/CD was never designed to address the subset problem, and 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.

4.4 Results

The transition probabilities between hypothesis states under a particular target language can be visualized as a state diagram. Unfortunately, the hypothesis space comprised some 10,000 grammars when the learner could start with any of the 720 M $\gg$ F grammars, and it would be quite difficult to visualize the whole system. It can be summarized, though. When the target language is English, the learner makes only four changes to his grammar before arriving at a hypothesis he does not escape from.

Figure 5 shows the rate of convergence for each of the 54 target languages, under the notions of extensional and superset success. All of the 720 M $\gg$ F grammars were allowed as initial states. Each line is a target language, and each iteration is a step of RIP/CD for one randomly-drawn overt form of the language. Convergence was quite fast, within roughly 30 iterations.

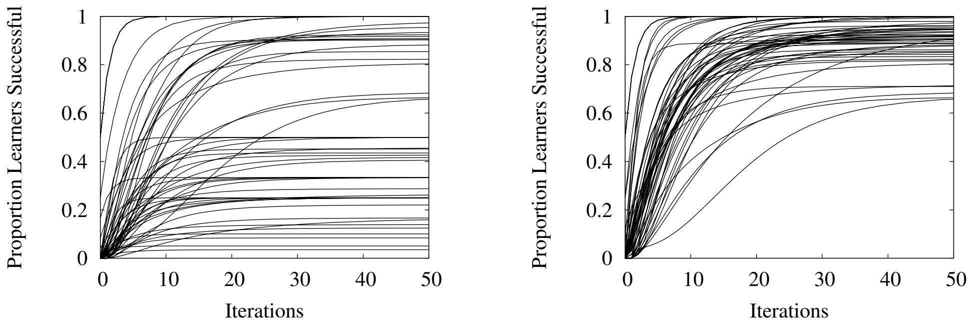

Figure 5: Convergence rates for the 54 target languages, under the notion of "extensional success" (left) and "superset success" (right). The initial states consisted of the 720 M $\gg$ F grammars.

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 6. With the 720 M $\gg$ F grammars as initial states here, the number of learnable languages under the extensional success criteria was 15, or 28%; under the superset success criteria, 36, or 67%. Neither of the other conceptions of UG—the bistratal or monostratal initial grammars—faired any better.

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

Figure 6: 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.

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 a sufficient number of tokens for the probability to settle on a value, is reported in Figure 7. Since these languages are attested or expected to be attested, we can determine which initial state sets and criteria are plausible. English is learned with the highest probability of only .89, for the 720 M $\gg$ F conception of UG

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 7: Probability of convergence for learners of three languages.

and after we grant leeway for the subset problem. Note that under the notion of extensional success, the learner never has a probability of 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.

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, but we could find no UG out of those we tried that supported learning all target languages.

5 Conclusion

The subset problem is an ongoing problem for RIP/CD and learnability theory generally. It is, however, not merely a matter of lacking negative evidence. RIP/CD generates its own negative evidence, but its negative evidence is imperfect. In particular, it fails to generate the kind of evidence needed to demote IDENTITY constraints.

The response to this problem in the literature has been to constrain Universal Grammar, restricting the initial grammar a RIP/CD learner can start with in the hopes that this sets him off on the right foot to learn whatever language he encounters. We showed through a simulation that the markedness-over-faithfulness restrictions on UG that have been described often in the literature are not sufficient if we expect a RIP/CD learner to be able to learn all possible languages in a factorial typology. 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 on a large scale any conception of UG will be sufficient for more than one target language.

The second freedom we have in evaluating RIP/CD is choosing how to evaluate successful learning. We used two methods, the first requiring that the learner match the target language's extension. This criterion is too strict if we hope to punt the subset problem to figure work: either narrowing down UG or revising the model of learning. So, we gave RIP/CD the benefit of the doubt and measured whether it at least leads the learner to a superset of the target language. It does not, not for all of the target languages at least, meaning that RIP/CD has problems other than the subset problem.[1]

Finally, the last degree of freedom we have is accepting that some languages in the factorial typology may be unlearnable (Boersma, 2003; Alderete, 2008). In that case, we can look only at the learnability results for attested languages in evaluating the learning procedure. We showed, however, that English does not fare well, and neither do two languages we considered based on attested languages. If Mandar′ with a probability of successful learning of at most .71 is attested, all three conceptions of UG and criteria for success that we considered would have to be abandoned.

This work provides some grounding for evaluating learning algorithms for Optimality Theory grammars, especially in relation to the subset problem, segmental alternations, and I–O constraints. These are 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.

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.

[1] In my master's thesis, I show that "Variationist EDCD," based on Boersma (2008), does appear to solve this problem, suggesting that the failure of RIP/CD under "superset success" here is due to the problems Boersma pointed out and claimed to have solved related to the refinement step.

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. 2008. 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.

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

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

Hayes, Bruce. 2001. Phonological acqusition in Optimality Theory: The early stages. In *Constraints in Phonological Acquisition*, ed. R. Kager, J. Pater, and W. Zonneveld, chapter 5, 158–203. Cambridge University Press.

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

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

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

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report RuCCS-TR-2, Rutgers Center for Cognitive Science.

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.

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.

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.

619 Williams Hall
University of Pennsylvania
Philadelphia, PA 19104
tauberer@ling.upenn.edu