Draft August 2014

Investigating the Efficiency of Parsing Strategies for the Gradual Learning Algorithm

Gaja Jarosz

1. Introduction

Theoretically motivated, constraint-based learning models are playing an increasingly pivotal role in phonological theory, including much research on variation (see Coetzee and Pater 2008b for a review), gradience and phonotactics (Coetzee and Pater 2008a; Hammond 2004; Hayes and Wilson 2008; Keller 2000; Martin 2011, and many more), acquisition modeling (see e.g. Boersma and Levelt 2000; Hayes 2004; Jarosz 2010; Jesney and Tessier 2011; Smolensky 1996; Tessier 2009), and inductive bias (Daland et al. 2011; Hayes and Londe 2006; Hayes et al. 2008; Wilson 2006, among others). The prominent role that constraint-based learning plays in the phonological literature is an important motivation for the current investigations into how such learners can be extended to successfully cope with structural ambiguity, such as that created by metrical feet, syllable structure, and other types of prosodic structure. Successful strategies for handling structural ambiguity have the potential to enrich and broaden the scope of research in each of the areas above by making accessible the modeling of empirical domains that rely on hidden structure or that interact with hidden structure. To give one concrete example, models capable of dealing with hidden structure need not make simplifying assumptions about the availability of syllable structure in the input (see discussion in Daland et al. 2011), but rather could simultaneously infer syllable structure and generalizations about well-formed syllables, clarifying the arguments that can be made about the role of the input as opposed to inductive biases.

There has been significant progress in recent years on the learnability of hidden structure in phonology, but a complete solution to this problem remains one of the most significant obstacles to realistic modeling of phonological learning. Much recent work has examined learnability in the constraint-based frameworks of Optimality Theory (OT; Prince and Smolensky 1993/2004), which relies on strict ranking of constraints, and in Harmonic Grammar (HG; Legendre et al. 1990, Smolensky & Legendre 2006), which relies on numerically weighted constraints. While the correctness and complexity of learning without hidden structure in OT and HG are relatively well understood (Tesar 1995; Pater 2008; Magri 2012; Bane et al 2010; Boersma and Pater 2008), very little is known about the correctness and complexity of constraint-based learners facing hidden structure, especially in the stochastic setting. Performance in the limit and efficiency are two critical properties of learning algorithms, and they have both played a role in the OT learnability literature from the very beginning. Correctness refers to a model's ability to consistently learn the target grammars, while efficiency refers to a model's capacity to converge sufficiently quickly on its solution. Both criteria are vital if learners are to have any hope of succeeding on the full complexity of the learning problem facing the language-learning child. Similarly, both are vital in practice in order to make sure that software implementing learning models returns reasonable grammars within a reasonable time frame.

A number of recent studies have investigated the success rates of OT and HG learning algorithms equipped with mechanisms for parsing structurally ambiguous data, with mixed results. For example, Boersma and Pater compared the success rates of OT and HG learners relying on one parsing strategy, Robust Interpretive Parsing (RIP; Tesar and Smolensky 1998). RIP assigns a parse to a structurally ambiguous form encountered during learning by selecting the fully structured candidate consistent with the input data that has the highest harmony according to the learner's current grammar. Boersma and Pater found that the algorithms' performance was promising for HG learners but quite low for classic OT and Stochastic OT learners, raising questions about the prospects of successful learning with hidden structure for OT learners. However, Jarosz (2013a) introduced two improved parsing strategies and showed that these new parsing strategies significantly improve success rates for both OT and HG and also eliminate the differences in performance between OT and HG learners. The present study investigates the efficiency of learning with RIP and the two parsing strategies proposed by Jarosz (2013a) – Resampling RIP (RRIP) and Expected Interpretive Parsing (EIP) – in the domain of metrical stress. The results indicate that RRIP and EIP improve not only end-state success rates, but also the speed of learning: learners equipped with the new parsing strategies converge on correct target grammars after fewer iterations of learning. The results suggest the new parsing strategies allow learners to extract information from the learning data more efficiently, improving the potential for learning to scale to the full problem of phonological learning. The discussion emphasizes the need for further work exploring a range of computational criteria in evaluating models of learning.

Before continuing it is important to emphasize that while the constraint-based modeling approach explored here comprises just one strand of a rich literature on the learnability of metrical structure, it differs from alternatives (Dresher 1999; Dresher and Kaye 1990; Goldsmith 1994; Heinz 2009) in providing a fully general approach to structural ambiguity that divorces the learning strategy from the substantive content of the phonological grammar. Like most other OT learning models, the RIP approach relies on the architecture of the grammar but is not tied to specific representations, constraints, or empirical domains (see Tesar 2004b for extensive discussion along these lines). This means that, while the present paper exemplifies the approach in a test system of metrical phonology, the proposed learning strategies can be applied to any case of structural ambiguity. In other words, the learning challenge undertaken by these approaches is a more general learnability problem that goes well beyond metrical phonology. Phonological theory posits various abstract representations to explain systematic regularities underlying surface patterns, including feet, syllable structure, moras, autosegments, and other prosodic structure. Metrical phonology is one domain in which such structure plays a central role, but it is important to keep in mind that further developments of constraint-based learning in this domain contribute to a deeper understanding of the challenges posed by phonological learning more generally.

2. Background

This paper examines the efficiency of several stochastic parsing strategies for the Gradual Learning Algorithm (GLA) for Stochastic OT (Boersma 1997, Boersma and Hayes 2001) and Noisy HG (Boersma and Pater to appear; Fischer 2005; Jäger 2007). Although the earliest of these parsing strategies, Robust Interpretive Parsing (RIP; Tesar and Smolensky 1998), was developed in the context of classic OT, it is in the stochastic setting that the differences between the parsing strategies become meaningful. This is because the parsing strategies differ only with respect to the probability with which different parses are selected. The present focus is to demonstrate the consequences these probabilistic choices have for efficiency of both OT and HG GLA learners. After briefly reviewing the GLA for the two frameworks, this section defines the three parsing strategies and summarizes previous work on their performance.

2.1 The OT and HG Gradual Learning Algorithms

This section briefly reviews the GLA for Stochastic OT (OT-GLA; Boersma 1997; Boersma and Hayes 2001) and Noisy HG (HG-GLA; Boersma and Pater, to appear; see also Rosenblatt 1958; Jäger 2007; Soderstrom et al 2006).

In both Stochastic OT (Boersma 1997; Boersma and Hayes 2001) and Noisy HG (Boersma and Pater to appear; Pater 2009; Pater to appear) constraints are associated with a value along a continuous scale. At evaluation time, random noise (sampled from a normal distribution centered around zero) is added to the value of each constraint independently, and the resulting relative ranking or weighting is used for optimization. In this way, Stochastic OT and Noisy HG define probability distributions over orderings and weighting of constraints, respectively, where constraints with similar values are more likely to re-rank or re-weight on different trials.

OT-GLA and HG-GLA are identical in all respects except for a minor difference in their update rules. OT-GLA makes use of the continuous scale by making small, repeated updates to the ranking values of constraints during learning, and HG-GLA does the same for weighting values of constraints. Since the GLA is an error-driven learning algorithm (Rosenblatt 1958; Tesar 1995; Tesar and Smolensky 1998; Wexler and Culicover 1980), adjustments to constraints' ranking and weighting values are triggered by errors. An error occurs when the learner's current grammar fails to generate the observed learning datum. In this case, the learner compares its output (the loser, L) to the learning datum (the winner, W) in order to determine how to adjust the grammar.

(1) Example Learning Trial

		5	4	2
	/dɔg/	*VOICE$_{\text{CODA}}$	*VOICE	IDENT[VOICE]
Winner (W)	a. [dɔg]	*	**	
	b. [dɔk]		*	*
Loser (L)	c. [tɔk]			**

Consider the example in (1), which shows the violations of three constraints relevant to obstruent voicing (*VOICE$_{\text{CODA}}$, *VOICE, and IDENT[VOICE]) for candidate outputs of 'dog'. Suppose the learner's current grammar associates the ranking or weighting values of 5, 4, and 2, respectively, with these three constraints, as shown in the top row of the tableau and that, for the purposes of illustration, noise is set to 0. When the learner observes the form [dɔg] (candidate a) as the output of /dɔg/, the learner first produces its own output for /dɔg/ and compares it with [dɔg]. In doing so, the learner makes an error since [tɔk] is the most harmonic output according to the current ranking or weighting. Thus, candidate (a) is designated the winner, while candidate (c) is designated the loser, and the learner must determine how to update the grammar by comparing the constraint violations of these two candidates.

(2) OT-GLA Update Rule
$\Delta r_i = \varepsilon \cdot \operatorname{sgn}(c_i(L) - c_i(W))$

(3) HG-GLA Update Rule
$\Delta w_i = \varepsilon \cdot (c_i(L) - c_i(W))$

The grammar update rules for OT-GLA and HG-GLA are shown in (2) and (3), respectively. As shown in (2), for each constraint *i,* OT learners compare the violations of the winner W and loser L under constraint *i*, $c_i(W)$ and $c_i(L)$, respectively. Constraint *i*'s ranking value is increased by *ε* (the learning rate or plasticity) when the loser has more violations than the winner and decreased by *ε* when the winner has more violations than the loser (the sgn function returns -1 for negative, 1 for positive, and 0 for zero). Thus, *ε* is added to the ranking values of winner-preferring constraints and subtracted from those of the loser-preferring constraints. As shown in (3), the update rule for HG is identical except that the plasticity is multiplied by the difference in the number of violations assigned to the loser and the winner.

Returning to the example, the constraint violations of the loser ($c_i(L)$) are (0, 0, 2), while the constraint violations of the winner ($c_i(W)$) are (1, 2, 0) for the constraints (*VOICE$_{\text{CODA}}$, *VOICE, IDENT[VOICE]). Subtracting the winner's violations from the loser's violations yields (-1, -2, 2). All that matters for the OT learner is which constraints are loser-preferring and which are winner-preferring so the OT learner will convert this vector to (-1, -1, 1) using the sgn function. A value of -1 indicates the constraint is loser-preferring since the winner has more violations on it, while 1 indicates the constraint is winner-preferring since the winner has fewer violations on that constraint. In this case, *VOICE$_{\text{CODA}}$ and *VOICE are loser-preferring, while IDENT[VOICE] is winner-preferring. The (-1, -1, 1) vector is multiplied by the learning rate *ε* and then added to the current ranking values. For example, if the learning rate were 0.1, then (-0.1, -0.1, 0.1) would be the grammar update, and this amount would be added to (5, 4, 2) to yield the new grammar (4.9, 3.9, 2.1). The HG update is identical except the sgn function is never used. This means the updates to the grammar depend on (-1, -2, 2) directly, and the weights would be updated to (4.9, 3.8, 2.2) with a learning rate of 0.1.

Thus, both OT-GLA and HG-GLA slightly decrease the ranking or weighting of all loser-preferring constraints and slightly increase the ranking or weighting of all winner-preferring constraints. The only difference is that OT learners move all constraints by the same amount, while HG learners make larger adjustments to constraints with a larger difference in violations between the winner and loser.

2.2 Hidden Structure and Parsing Strategies for the GLA

The learning strategy just described assumes the learner is provided with full structural descriptions of the learning data. This means the learner has access to hidden structure such as footing and syllabification as well as underlying representations, which are not available to the human learner. Access to such hidden structure is crucial for identifying the violations incurred by the winners, which, as just discussed, are crucial for calculating the update. This is because every overt form is structurally ambiguous and corresponds to numerous distinct candidates, each with different hidden structures and therefore distinct constraint violations. For example, a tri-syllabic word with medial stress such as [tɛˈlɛfɔn] has at least two parses, one with an initial iambic foot [(tɛˈlɛ)fɔn], and one with a final trochaic foot [tɛ(ˈlɛfɔn)]. The constraint violations of constraints such as IAMBIC, TROCHAIC, ALLFEETLEFT, and ALLFEETRIGHT for this form depend entirely on the parse. The iambic parse violates ALLFEETRIGHT and TROCHAIC, while the trochaic parse violates the opposite constraints: ALLFEETLEFT and IAMBIC. Without access to the parse the learner has no way of knowing which constraints prefer the loser and which prefer the winner. This is the challenge that hidden structure poses for the error-driven learner. In essence, hidden structure obscures the constraint violations incurred by the learning data, which in turn obscures the grammar updates that must be made in response to errors.

Within OT and related constraint-based frameworks, there is an extensive body of work addressing this idealization about the learning task in pursuit of learning models with some mechanism for handling hidden structure (Akers 2011; Alderete et al. 2005; Jarosz 2006a; Jarosz 2006b; Jarosz 2013a; Jarosz 2013b; Merchant 2008; Merchant and Tesar 2008; Prince and Smolensky 2004; Tesar 1997; Tesar 2004a; Tesar 2004b; Tesar 2006a; Tesar 2006b; Tesar 2008; Tesar 2009; Tesar et al. 2003). RIP and the other parsing approaches to structural ambiguity investigated here form one prominent strand of learnability research from the constraint-based perspective. These parsing strategies have been applied to both strict and probabilistic variants of both ranking and weighting frameworks (Apoussidou 2007; Apoussidou and Boersma 2003; Biró to appear; Boersma 2003; Boersma and Pater to appear; Jarosz 2013a; Tesar and Smolensky 1998; Tesar and Smolensky 2000), and RIP has also been adapted to the learning of hidden lexical representations (Apoussidou 2006; Apoussidou 2007).

One appealing aspect of these approaches is that they can endow widely used learning algorithms such as the GLA with the ability to handle hidden structure without fundamentally altering the underlying algorithms. The algorithms retain their sensitivity to frequency and their simple, online processing of the learning data. The only modification is that the parsing strategies are used to assign structure to the learning data as each form is processed. The algorithms are otherwise unchanged. Since stochastic variants of OT and HG are playing an increasingly prominent role in phonology, it is important to continue to develop and study the extensions of learning algorithms for these frameworks that allow these models to be applied to richer learning data in a more realistic way.

In order to perform error-driven learning in the presence of structural ambiguity, Tesar and Smolensky (1998) proposed Robust Interpretive Parsing (RIP), which provides an educated guess, based on the current constraint ranking, about the structure of the observed datum. RIP was later applied to OT-GLA (Apoussidou 2007; Apoussidou and Boersma 2003; Boersma 2003) and HG-GLA (Boersma and Pater to appear), and it is these stochastic learning algorithms that are the focus presently. RIP uses the learner's current hierarchy to select the most harmonic candidate among the structural descriptions consistent with an overt form. That is, for a given learning datum, RIP uses standard OT or HG optimization but limits candidates to those that share the learning datum's overt form, thereby selecting the most harmonic among the possible structural descriptions, or parses, of the overt form according to the current grammar.

The full RIP/GLA learning procedure for both OT and HG works as follows. For each learning datum, the learner randomly samples a ranking or weighting from the current stochastic grammar, just as in regular GLA. The learner then uses that ranking or weighting to perform Robust Interpretive Parsing on the learning datum, identifying the winner. The learner uses the same ranking or weighting to generate its own output for that learning datum and compares it to the winner. This output is called the loser when it differs from the learning datum. Given the winner and the loser, the grammar update, comparing the violations of the parsed winner to that of the fully structured loser, can proceed as usual according to the update rules discussed above.

An example for the structurally ambiguous [tɛˈlɛfɔn] is shown in (4). According to RIP, the learner must first parse the ambiguous [tɛˈlɛfɔn] and then compare the parse to its own output to determine whether an error has been made. The parse is found by considering only those candidates that have the same overt form as the learning datum. In this case, this includes the two candidates with medial stress, (unshaded) candidates (b) and (c). Using the ranking shown in the tableau, the OT learner will select candidate (c) as the harmonic structure of the two, and (c) will therefore be designated as the parse. The RIP learner then uses the same ranking to compute its own output, which in this case will be candidate (d) [tɛ(lɛˈfɔn)]. These are not identical so (c) is designated the winner, (d) the loser, and the grammar update is calculated as usual according to the update rules discussed earlier. In HG, these steps work identically, except weighting rather than ranking is used for optimization of both the parse and the output.

(4) Robust Interpretive Parsing for [tɛˈlɛfɔn] (example from Jarosz 2013a)

	/tɛlɛfɔn/	AFR	IAMB	TROCH	AFL
	a. (ˈtɛlɛ)fɔn	*	*		
	b. (tɛˈlɛ)fɔn	*		*	
RIP parse (Winner)	c. tɛ(ˈlɛfɔn)		*		*
(Loser)	d. tɛ(lɛˈfɔn)			*	*

After identifying two problems with the above formulation of RIP for the GLA, Jarosz (2013a) proposed two alternative parsing strategies: Resampling Robust Interpretive Parsing (RRIP) and Expected Interpretive Parsing (EIP). The reader is referred to that paper for extensive discussion of all three parsing strategies and their properties. For present purposes, what is crucial are the differences between the three parsing strategies.

There are two differences between RIP/GLA and RRIP/GLA. The first difference is that in RRIP/GLA the learner ignores hidden structure when comparing its output to the learning datum, evaluating only whether the forms match in their overt material. Since the structure of the learning datum is irrelevant for the comparison, parsing is necessary only for purposes of calculating the grammar update when an error has occurred and need not be performed up front. The second difference is that RRIP/GLA uses a new sample from the current stochastic grammar to calculate the parse rather than using the same sample that is used to generate the learner's own output, as in RIP. That is, the learner resamples from the grammar before applying RIP to the learning datum. This ensures the RRIP learner has a chance to select a ranking or weighting from the current grammar that does not generate the same error. As a consequence, the learner falls back on its current probabilistic grammatical knowledge to select a ranking or weighting it has confidence in and uses that ranking or weighting for parsing. In contrast, the original RIP procedure is doomed to parse using a ranking or weighting that led to an output error, a behavior Jarosz showed impacts performance negatively because it creates internal inconsistency for the learner.

To illustrate the difference between RIP and RRIP, consider the modified example in (5). Suppose that the stochastic grammar has IAMBIC and TROCHAIC variably ranked at the top of the hierarchy so that each relative ranking of these two constraints is equally likely but ALLFEETRIGHT is ranked above ALLFEETLEFT with very high probability. Ranking values that would have this effect would be (30, 30, 20, 10), for example. Suppose further that the target language has penultimate stress, requiring right-aligned trochees. This means the learner's current grammar is close to correct: all that is needed is for TROCHAIC to increase its ranking relative to IAMBIC. The learner will still make errors on the target form [tɛˈlɛfɔn], producing candidate (d) with final stress whenever the selected ranking has IAMBIC ranked above TROCHAIC. In this situation, RIP will always select the wrong parse for this target language, candidate (b), because that is the parse that is optimal under that relative ranking IAMBIC ≫ TROCHAIC, and RIP uses the same ranking for parsing as the one that triggered the error. When candidate (b) is incorrectly selected as the parse and used to calculate the grammar update, the learner is led astray into adjusting the relative ranking of ALLFEETLEFT and ALLFEETRIGHT, rather than the footform constraints, as required for successful learning. In contrast, RRIP selects a new ranking from this grammar for parsing when an error occurs, and has a chance of selecting a ranking with TROCHAIC ≫ IAMBIC. In this example, RRIP therefore has a 50% chance of selecting the correct parse, candidate (c), and making the correct update for this target language, increasing the ranking values of TROCHAIC and decreasing those of IAMBIC.

(5) Robust Interpretive Parsing for [tɛˈlɛfɔn]

	/tɛlɛfɔn/	IAMB	TROCH	AFR	AFL
	a. (ˈtɛlɛ)fɔn	*		*	
RIP parse (Winner)	b. (tɛˈlɛ)fɔn		*	*	
	c. tɛ(ˈlɛfɔn)	*			*
(Loser)	d. tɛ(lɛˈfɔn)		*		*

The second strategy proposed by Jarosz, EIP/GLA, is nearly identical to RRIP/GLA except that the two-step parsing process of RRIP (resampling followed by RIP) is replaced with sampling a parse from the conditional probability of the parse given the current grammar and the learning datum. This means a parse is selected probabilistically in proportion to the likelihood with which the grammar generates that structure for this overt form. Jarosz implements this strategy by sampling from the production grammar until a matching overt form is found and using that first match as the parse. EIP is the only parsing strategy that selects parses from a distribution that is compatible with the learner's production grammar. In this way EIP takes full advantage of the rich information available in the learner's current grammar hypothesis.

The effect of this modification can also be illustrated with the example in (5). The learner's current grammar can only produce two outputs with substantial probability, candidate (c) and candidate (d), depending on the (variable) relative ranking of IAMBIC and TROCHAIC. This is because the grammar already encodes a preference for rightward alignment of feet (AllFeetRight ≫ AllFeetLeft), and only right-aligned feet are generated. EIP parses consistently with this grammatical knowledge: when a parse has to be assigned to ambiguous [tɛˈlɛfɔn], EIP consistently returns candidate (c) (tɛ(ˈlɛfɔn)), since that is the only matching output it generates using its production grammar, which only generates right-aligned feet. Consequently, by relying fully on the stochastic grammatical knowledge the learner has already accumulated, EIP selects the correct parse and makes the correct grammatical updates more reliably than either RRIP or RIP.

To summarize, all three parsing strategies provide a way to probabilistically assign structure to ambiguous learning data, but they differ in how information from the current grammar is utilized to do so.

2.3 Summary of Previous Results and Simulations

The simulations presented in this paper rely on the same metrical phonology test set used to evaluate RIP, RRIP and EIP for both OT and HG in previous work (Boersma 2003; Boersma and Pater to appear; Jarosz 2013a, 2013b; Tesar and Smolensky 2000). This allows for direct comparison with previously reported results. This section presents that test set and reviews the previous results.

This test set, first defined and examined by Tesar and Smolensky (2000), consists of 124 constructed languages that can be modeled by the set of twelve metrical structure constraints shown in (6). Most of these constraints are well known from the literature, with origins in the early OT literature (McCarthy and Prince 1993; Prince and Smolensky 2004) and pre-OT metrical phonology (Hayes 1995; Liberman and Prince 1977; Prince 1990). One exception is the non-standard formulation of the constraint favoring trochees: Foot-NonFinal (Tesar 2000). The interaction of these twelve constraints produces a complex system, inspired by natural language stress systems, capable of describing a range of diverse metrical phenomena. The test system has the crucial property of generating structural ambiguity – overt stress patterns in this system are consistent with multiple structural descriptions, and successful learning requires disentangling interdependent and ambiguous requirements made by the individual learning data. Tesar and Smolensky selected 124 languages from the factorial typology generated by this constraint set to represent a wide range of metrical phenomena.

(6) Constraints (Tesar and Smolensky 2000)

FootBin	Each foot must be either bimoraic or disyllabic
Parse	Each syllable must be footed
Iambic	The final syllable of a foot must be the head
Foot-Nonfinal	A head syllable must not be final in its foot
Nonfinal	The final syllable of a word must not be footed
WSP	Each heavy syllable must be stressed
Word-Foot-Right	Align right edge of the word with a foot
Word-Foot-Left	Align left edge of the word with a foot
Main-Right	Align head foot with right edge of the word
Main-Left	Align head foot with left edge of the word
All-Feet-Right	Align each foot with right edge of the word
All-Feet-Left	Align each foot with left edge of the word

Each language in the system is defined by a set of surface stress patterns for sixty-two words that can be generated from this constraint set. Words are sequences of light (L) or heavy (H) syllables ranging in length between two and seven syllables (e.g. [H L H L]). Each word is associated with a surface stress pattern (e.g. [H1 L0 H2 L0]), indicating for each syllable whether it has primary stress (1), secondary stress (2), or no stress (0). Any given ranking or weighting of the constraints assigns a particular foot structure and pattern of stress (e.g. [(H1 L0) (H2) L0]). Indeed, it is the footing that underlies the systematic stress patterns in the system. The learner, however, is exposed only to the overt stress patterns (e.g. [H1 L0 H2 L0]) and must infer a ranking or weighting of constraints (and an associated footing) capable of generating the observed surface stress patterns. The learner is considered successful when it has acquired a grammar that is consistent with all the learning data it is exposed to, that is, when it assigns the correct surface stress patterns to all the words of the language.

(7) End State Success Rates (noise) of GLA for OT and HG Depending on Parsing Strategy

	OT	HG
RIP (Boersma & Pater 2013)	58.95 (2)	88.63 (2)
RIP (Jarosz 2013a)	56.13 (2)	91.05 (4)
RRIP (Jarosz 2013a)	82.58 (2)	92.42 (4)
EIP (Jarosz 2013a)	93.95 (2)	94.19 (4)

Table (7) summarizes the success rates (proportion of languages correctly learned) for the three parsing strategies applied to the GLA for OT and the GLA for HG reported in previous work. All the simulations in Boersma & Pater (2013) and Jarosz (2013a) are averaged success rates over 10 separate runs for each of the 124 languages. On each run the algorithms were allotted a maximum of 1,000,000 iterations, where each iteration corresponds to the processing of one overt form. The algorithm is considered successful for a particular language if, when evaluation noise is set to zero, the final grammar correctly generates the surface stress patterns for all words in the language. For all simulations above, the initial ranking or weighting values were set to 10 and the plasticity to 0.1. Boersma and Pater used a noise value of 2 for both models, while Jarosz explored several additional parameter settings and found that the HG models were more successful with a noise setting of 4. The results shown for HG are for this noise value.

As the results from both studies indicate, when RIP is used as a parsing strategy, the HG-GLA learners massively outperform the OT-GLA learners. Jarosz (2013a) showed, however, that RIP is a suboptimal procedure that disproportionately affects the OT learners. The RRIP and EIP parsing strategies make increasingly better use of the stochastic grammatical information available to the learner, as evidenced by the improvement between RIP and RRIP for both OT and HG and between RRIP and EIP for both OT and HG. The improvement for OT-GLA is much more substantial, however (for discussion see Jarosz 2013a), with end state success rates for EIP/OT-GLA and EIP/HG-GLA being comparable. These results indicate that RRIP and EIP not only improve success rates for both weighting and ranking frameworks, but they also level the playing field between HG and OT in this context. When these improved parsing strategies are used, there is no longer an advantage for HG.

The focus of the present investigation is to determine whether the parsing strategies that improve end state success rates also make the learners more effective at processing incoming information, leading to quicker convergence on the target grammars.

3. Considerations of Computational Effort

In the previous work discussed above, all the learning algorithms were allowed the same number of iterations, 1,000,000, to process the data, and their success rates given this maximum number was calculated. However, the relative success rate of learning algorithms after a fixed number of iterations is only one facet of an algorithm's performance. Another important consideration, and one that has played a prominent role in the literature on learnability within classic OT from the very beginning, is one of algorithmic efficiency. Starting with the earliest work on OT learnability, the learnability results for the Constraint Demotion family of learning algorithms included proofs not only of their correctness but also of their data complexity (Tesar 1995; Tesar and Smolensky 1998). In particular, Tesar and Smolensky showed that, with access to full structural descriptions, both Error-Driven Constraint Demotion (EDCD) and Recursive Constraint Demotion (two variants of constraint demotion for classic OT) have efficient data complexity. For *n* constraints, both algorithms require no more than $n(n\text{-}1)$ errors to converge on a target ranking for the language, assuming the data are consistent with a total ranking of constraints.

Considerations of data complexity are important for evaluating the feasibility of a learning algorithm. When the full complexity of natural language is considered, the hypothesis space of possible target languages is enormous (see e.g. Tesar 2006a for extensive discussion along these lines). Thus, it is generally assumed that exhaustively or randomly searching this space is not a feasible strategy for language learning. It follows, then, that any learning algorithm that requires more data to learn than exhaustive or random search of all possible grammars is also not feasible. On small, constructed test systems, the success rate of such a learner might be high if it is provided with sufficient iterations, where each iteration corresponds to the processing of one datum. However, if the learner does not learn more efficiently than random search, it is doomed to fail on larger, more realistic learning problems. Therefore, comparing the amount of data needed for successful learning against a baseline algorithm such as random search provides a simple test of the algorithm's potential to scale up. If the algorithm requires more data for successful learning than random search, this does not bode well for its prospects on more realistic learning problems.

While data complexity of EDCD and RCD is well understood, little attention has been given to the data complexity of the stochastic incarnations of RIP: RIP/OT-GLA and RIP/HG-GLA. In fact, recent work has raised questions about the interpretation of end state performance results on tests sets such as the one explored here and in previous evaluations of RIP/GLA. Jarosz (2013b) explores several learning algorithms for a probabilistic version of OT and compares their performance to a random baseline on the same test set used in the present work. Although the algorithms' performance appears promising, Jarosz finds that the random baseline, given 1,000,000 data forms to process, has a 100% success rate on this test system[1]. The main import of this result for the present work is that end state success rates are only part of the picture, and the hypothesis space of this test system is not as challenging as one might expect based on the number of constraints. It is therefore vital to consider these results in the context of some baseline. A successful algorithm must not only learn the target languages in its hypothesis space, it must also be capable of converging on the target language in a reasonable time. This is an important question because the high success rates of RIP/HG-GLA and the RRIP and EIP versions of OT-GLA are meaningful only if the learning strategies are feasible and promise to scale better than the baseline. After all, performance in the limit is irrelevant if learners cannot realistically be exposed to enough data to ever reach this level of performance.

In consideration of these important concerns, the following sections examine the learning curves for each of the three parsing strategies and explicitly compare the learning curves of the learning algorithms against a random baseline.

3.1 The Random Baseline

Following Jarosz (2013b), the random baseline explored here relies on a very simple learning strategy: random search. In essence, the random baseline selects a ranking or weighting at random and uses it to process each incoming datum. If the current (random) grammar generates the stress pattern for the datum, the baseline does nothing and moves on to the next datum; if not, the baseline simply randomly selects another ranking or weighting before moving to the next datum. Essentially, the random baseline's learning strategy entails picking rankings or weightings at random until errors are no longer produced.

As Jarosz (2013b) showed, this random OT baseline achieves perfect performance on this test set when given 1,000,000 iterations. More specifically, given a maximum of 1,000,000 iterations (where an iteration corresponds to the processing of one datum, as before), the random OT baseline learns all the languages on all runs (with 10 runs for each language). It is not surprising that the random baseline eventually achieves a 100% success rate; a random baseline will eventually find the target grammar in any finite hypothesis space, such as the space of total rankings of twelve constraints. More surprising is the fact that the random baseline consistently succeeds within only 1,000,000 iterations even though the number of total rankings, 12!, is almost 500 times larger than this. As Jarosz explains, this discrepancy highlights the extent to which different rankings in this test system are weakly equivalent. An algorithm is successful if it succeeds in generating the stress patterns in the learning data, and there are usually multiple distinct rankings that generate any given stress pattern. In fact, the success of the random OT baseline demonstrates that on average there are hundreds of rankings consistent with each language in this test system.

From the perspective of an HG learner, however, the hypothesis space created by these constraints is much larger. Previous work has shown that constraint weighting is a more powerful form of constraint interaction than constraint ranking, yielding gang and

[1] In fact, some of the algorithms Jarosz introduces also achieve 100% success rate, but they do so more slowly than the random baseline, and as Jarosz discusses, are therefore not viable learning algorithms.

other cumulative effects not possible in OT (Bane and Riggle to appear; Goldwater and Johnson 2003; Jäger 2007; Keller 2000; Pater 2009; Potts et al. 2010; Prince and Smolensky 2004; Smolensky and Legendre 2006). The additional power of weighting means that the same set of constraints can generate a larger predicted typology under weighting than under ranking. For example, a set of five simple syllable structure constraints predicts just 12 distinct languages in OT but 23 languages in HG (Bane and Riggle to appear). Furthermore, gradient alignment constraints (McCarthy and Prince 1993), such as those used extensively in the stress system examined here are particularly prone to non-OT-like interactions (Bane and Riggle to appear; Legendre, Sorace and Smolensky 2006; Pater 2009). Indeed, Bane and Riggle also examined a stress system making use of twelve alignment constraints (Gordon 2002) and found that while OT predicted a total of 152 distinct stress systems, HG predicted 36,846 distinct systems when word lengths were restricted to eight syllables (otherwise, the typology was infinite).

This means that the efficiency of HG learners on this set of languages cannot be directly compared to the efficiency of OT learners, which are navigating a smaller hypothesis space. In general, the amount of data needed for successful learning depends on the size of the hypothesis space, that is, the size of the predicted typology in the present context. As the typology gets larger, the target language takes up an increasingly smaller proportion of the hypothesis space, making it harder for the learning algorithm to find the target language. On average, therefore, a larger typology makes it more difficult for a learning algorithm to learn any particular target language in that typology. Therefore, it is necessary to construct a separate random baseline for HG learners and evaluate the HG models against it. A random search in HG space will reflect the magnitude of the hypothesis space with which the HG learners are grappling. Randomly searching weightings is somewhat more complex than uniformly selecting a ranking from the finite set of possible rankings. In Stochastic OT only the *relative* ranking values matter so that ranking values of 110 and 102 are equivalent to 10 and 2, respectively. This is not so in HG. In HG there are multiplicative interactions where multiple violations of a lower weighted constraint can overpower a higher weighted constraint, and these interactions depend on the absolute weights. For example, two violations of a constraint weighted at 102 can overpower one violation of a constraint weighted at 110, but the same is not true for the equivalently separated weights of 10 and 2. In order to address this richer range of weightings in HG, a number of random HG baselines was considered here, and the best was selected for comparison to the models. Specifically, the random baselines select weightings randomly according to an HG grammar that weights all constraints equally. For example, one random HG baseline selects weightings from a grammar whose constraints are all tied at 10 using a noise value of 8. There are infinitely many such tied HG grammars since they may vary in the weighting value of the constraints and in the noise, and these parameters affect the distribution of weightings that are randomly selected. For purposes of these evaluations the weighting value and noise combination that resulted in the best performance was used, and this was the HG baseline with all constraints weighted at 10 and noise set to 16.

The previous simulation results with the random OT baseline put the performance of the RIP, RRIP, and EIP variants of OT-GLA in perspective. These results show that this test set, despite its apparent complexity, involves a rather small hypothesis space of weakly equivalent rankings. This hypothesis space is small enough that random OT search can consistently navigate it successfully within 1,000,000 iterations. Although RIP, RRIP, and EIP for OT-GLA do not achieve 100% success given the same number of iterations, this does not necessarily mean that these algorithms are doomed. It does mean, however, that measuring success rates after 1,000,000 iterations cannot be the only consideration for evaluating performance on this test set. It is important to consider the success rate of a baseline such as random search compared to RIP, RRIP and EIP over the course of learning for OT learners. While the space of hypotheses is much larger for HG learners, their performance must likewise be evaluated against a baseline for HG.

Examining the learning curve will indicate how successful random search is if it is given fewer iterations for learning. Similarly, it will indicate how the success rates of RIP, RRIP, and EIP depend on the number of iterations spent processing data. Based on the results already presented, it is clear that no OT learning algorithm can ever beat the success rate of the OT baseline in the limit – or even 1,000,000 for this test set – since the success rate of the OT baseline is 100%, which cannot be beat. However, how does the success rate of RIP, RRIP, EIP, and the random baselines depend on the amount of learning data processed? How quickly (or slowly) does the random baseline overtake OT-GLA and how do the HG-GLA learners fare relative to their baseline? As discussed above, a learning algorithm's potential to scale up to larger problems can be measured by how much more quickly it learns than the random baseline before it plateaus. The following sections present the results of comparisons with the random baselines, considering two different ways of evaluating successful learning, in turn.

3.2 Results: Proportion of Languages Correctly Learned

To answer these questions, the intermediate grammars of all the learners discussed above – including the two baselines – were inspected periodically to check for successful learning. The average success rates of each of the algorithms was calculated at each inspection point by using the learner's current grammar with noise removed and generating the learner's output for each of the 62 forms in the data. The first evaluation defines successful learning in the same way as in previous work: the learner is successful if it correctly generates the stress patterns for all the forms in the language. Thus, the average success rates are the average number of languages correctly learned at each inspection point. The results of this inspection were not available to the learning algorithms themselves, which were set up exactly as described in the preceding sections. The average success rate at each inspection point was also calculated for the random baselines. For all algorithms and the baselines, inspection points identify the total number of data forms processed so far (iterations), regardless of whether errors were produced. This yields a learning curve indicating the relationship between average success rate and the number of iterations.

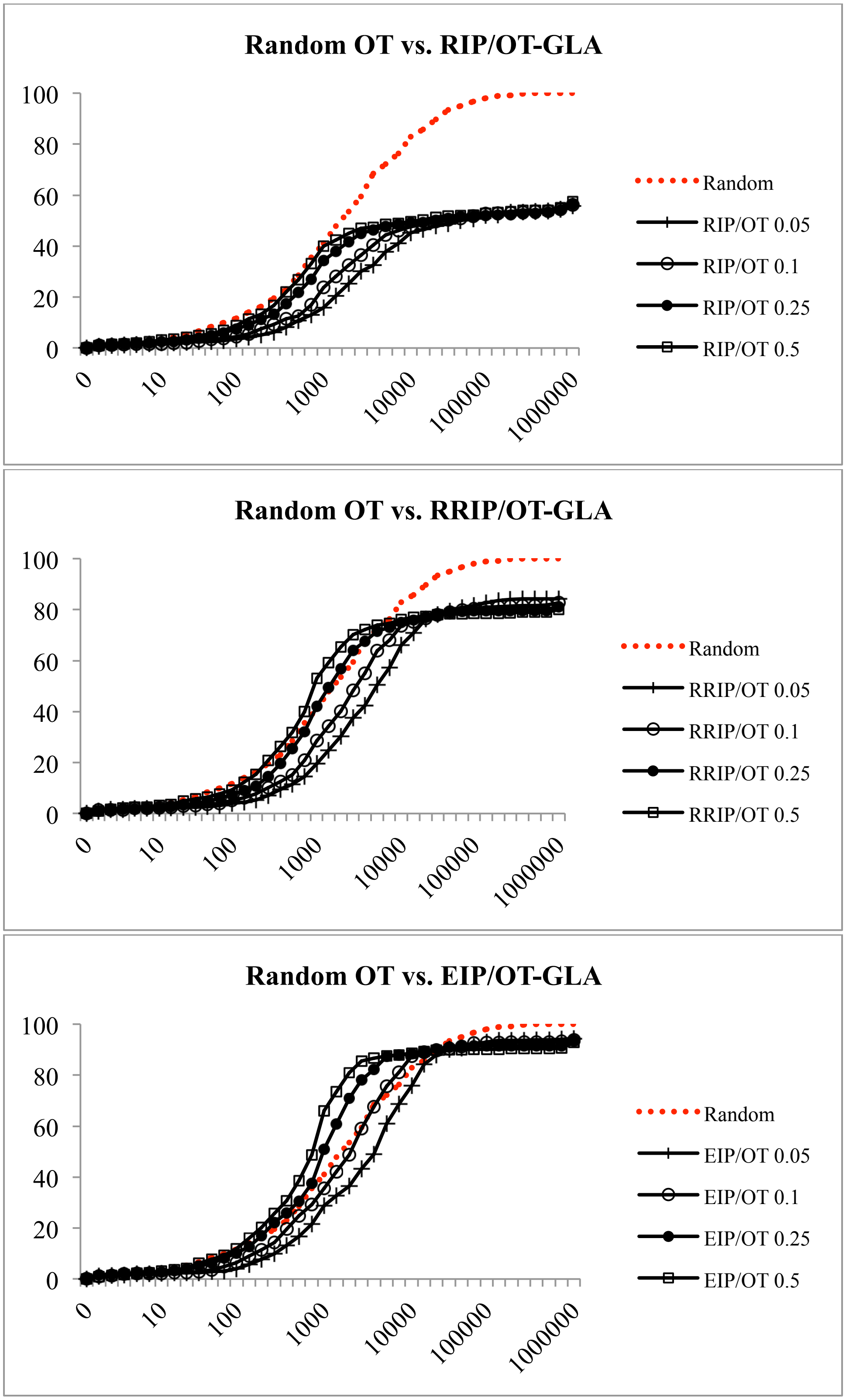
Figure 1 and Figure 2 below compare the learning curves of the respective baselines to each of the three parsing strategies for both OT-GLA and HG-GLA at various parameter settings. The results already established indicate that toward the end of the learning, the random OT baseline will reach 100%, and the other learning algorithms will reach the end-state success rates reported in previous work. The purpose of these examinations is to determine the shape of these curves during earlier stages of learning.

In all figures the iterations (on the x-axis) are shown on a log scale, making it easier to compare performance differences at the earlier stages of learning.

Figure 1 shows the results for the three OT-GLA algorithms, RIP/OT-GLA, RRIP/OT-GLA, and EIP/OT-GLA, at various parameter settings. These curves reveal several important properties about the performance of these algorithms. First, the curves for RIP fall almost entirely below the curve of the random baseline. Although higher learning rates allow the algorithm to learn more quickly initially, even the learners with higher learning rates have a limited window of opportunity to rise above the baseline since they plateau at such a low success rate. As discussed by Jarosz (2013a) increasing the learning rates further leads to even poorer success rates at the end of learning. A second observation is that the curves for RRIP and EIP are successively steeper. RRIP clearly crosses above the baseline curve, and EIP performs even better. This shows that the improvements in parsing mechanisms afforded by RRIP and EIP affect not only the algorithms' performance in the limit but also their efficiency. These improvements are cumulative, with EIP performing better than RRIP. Overall, these results indicate that the RRIP and EIP algorithms improve the GLA's potential to scale up. Furthermore, if the measure of success used here is taken at face value, these results suggest that RIP/OT-GLA has little hope of scaling up to larger problems. In particular, the results of these simulations indicate that the RIP/OT-GLA learners are slow to settle on target grammars that succeed in correctly generating all the overt forms in the learning data. This success metric is discussed further after reviewing the results for HG-GLA.

The results RIP, RRIP, and EIP for HG-GLA at various parameter settings are shown in Figure 2. The fact that there is only a subtle difference between RIP, RRIP, and EIP is not particularly surprising since the end state success rates of these three algorithms reported by Jarosz (2013a) varied relatively little. More informative is the fact that all the HG-GLA models are above the HG baseline throughout the course of learning. Additionally, the results indicate that the choice of parameters used for HG-GLA learners affects the learning curves in complex ways. For example, the heights of the curves don't consistently correlate with higher or lower learning rates. This reiterates Jarosz's (2013a) point that further work is needed investigating the consequences, both in terms of success rates and efficiency, of various parameter settings used for the HG-GLA learners.

In sum, this section has shown that the new parsing strategies, RRIP and EIP, improve the performance of the GLA from the perspective of data complexity. RRIP allows the GLA to learn more quickly from the data than RIP, and EIP allows the GLA to learn more quickly than RRIP. These results are dramatic for OT-GLA and more subtle for HG-GLA. These results make sense considering the earlier discussion about RIP's failure to make full use of the learner's grammatical knowledge. Better reliance on the learner's stochastic grammar by RRIP and EIP leads to faster learning. By explicitly considering performance of baseline, this section also demonstrates that the choice of parsing strategy has potentially dramatic implications for RIP/OT-GLA. Because RIP/OT-GLA performs consistently below the random baseline, this strongly suggests it is not likely to cope well with larger and more realistic learning problems. However, this pessimism is only warranted to the extent that the success metric employed in this section reflects reasonable criteria for stochastic learning models. The next section considers this question in more depth.



Random OT vs. RIP/OT-GLA

Random OT vs. RRIP/OT-GLA

Random OT vs. EIP/OT-GLA

Figure 1 – Success Rate of Random OT Baseline vs. OT-GLA Over Time

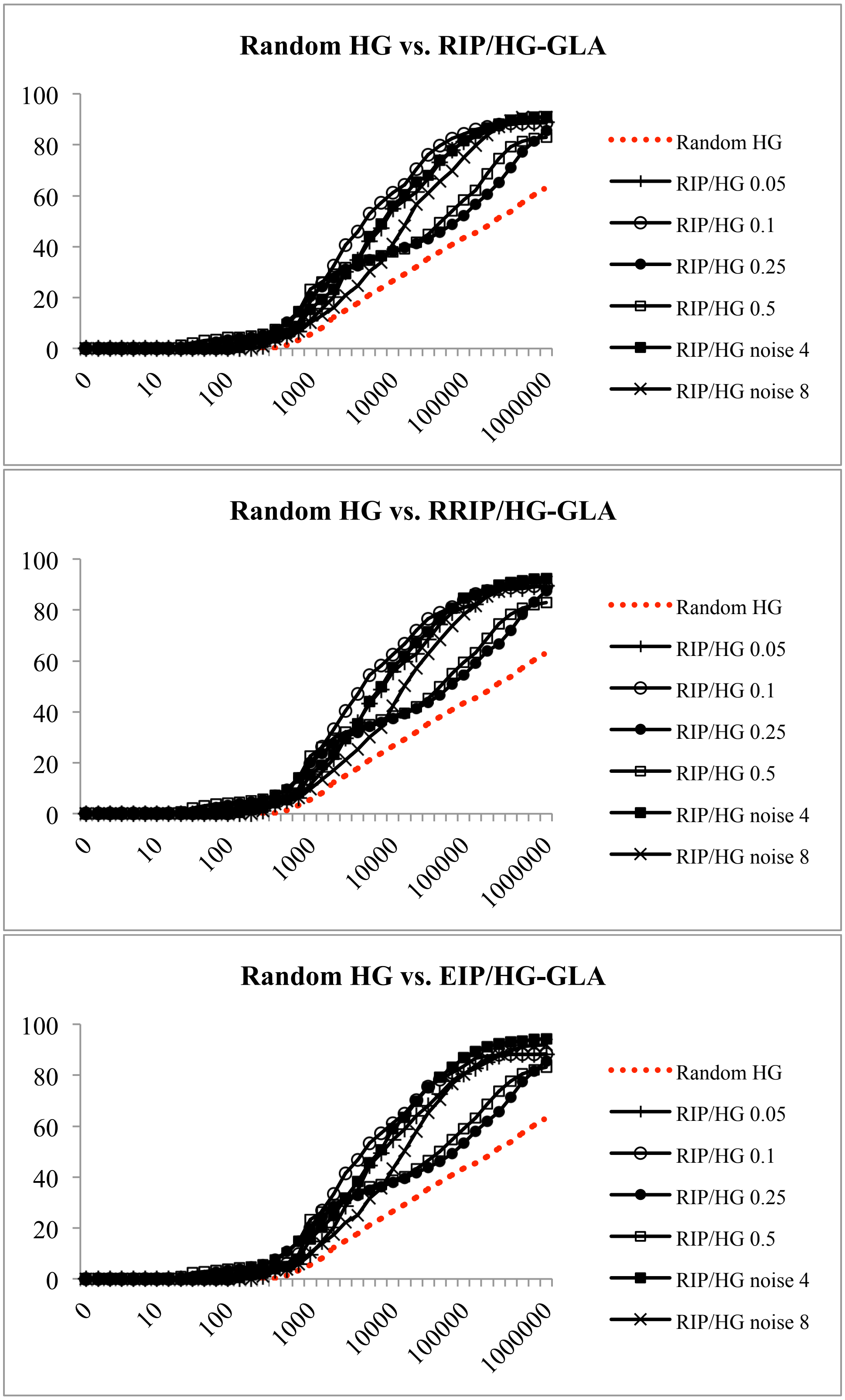


Figure 2 – Success Rate of Random HG Baseline vs. HG-GLA Over Time

3.3 Results: Proportion of Word Classes Correctly Learned

The previous sections have examined end-state success rates and learning curves using a standard measure of performance in the OT literature: the proportion of languages correctly learned. This is a very strict metric that gives no partial credit, and for the problem of learning categorical languages, there are reasons to think it is an appropriate measure. In the test system examined here, getting one overt form wrong means getting the stress pattern for an entire class of words wrong. For example, if the model incorrectly learns the stress pattern for LHHL, it has arguably failed to learn the target grammar as it assigns the wrong stress contour to all sequences of LHHL. Assuming such a grammar instantiates a kind of stress system that learners must be capable of learning, the measure of success should severely penalize a learner that fails to learn it. A major goal of much work in OT is to define a universal set of constraints that, under permutation, generates a set of possible languages. From this perspective, a constraint set defines the set of languages that should be learnable, and a model should be penalized for getting any of these languages wrong. In other words, models (or perhaps more accurately, modelers) should aspire to perfect performance on all data.

However, the focus of this paper is on probabilistic models, a major advantage of which is the capacity to learn languages with variability. If learners are to be evaluated on their capacity to learn variable as well as non-variable target languages, this effectively changes the learning problem. Viewed from this perspective, the metric used above is an unattainable ideal since models cannot reasonably be expected to perfectly reproduce freely variable learning data. Furthermore, the learning data is only a finite sample, not the true distribution of possible forms in the language so requiring learners to perfectly match proportions in the data does not make sense either. A more lenient evaluation metric is needed, one that gives partial credit for getting closer to the right answer. One simple way to define a more lenient metric is to use per word accuracy rather than per language accuracy. Such a metric indicates the proportion of words in the language generated correctly, giving partial credit for partially correct grammars. This is still not an ideal way of measuring success for this learning problem since it can fail to detect small but systematic divergences from the target grammar. However, it does provide additional valuable information, contributing to a more comprehensive understanding of the performance of the models.

Before examining the learning curves, the way in which the new success metric evaluates end-state success of the six learning models is summarized in Table (8). These percentages correspond to the proportion of times a word form is produced with the correct stress contour, averaged over all word forms in each language, all languages in the system, and ten separate runs for each language. For each model, these results show the end-state accuracy for the learning rate and noise setting that yielded the steepest learning curve in the previous section. Not surprisingly, the numbers are overall much higher as compared to the per language success rates. This indicates that in general, when the learners fail to perfectly reproduce the stress patterns for a language, they nonetheless tend to come close, generating the correct stress pattern for most words most of the time. Additionally, these results confirm the major effects established earlier. Performance of RIP/OT-GLA is lower than RRIP/OT-GLA, which is lower than EIP/OT-GLA, and these effects are both highly significant ($p < .001$, Welch two sample t-test). This confirms the earlier findings that each of the parsing strategies cumulatively improves over performance of OT-GLA with RIP. Also highly significant are the differences between EIP and each of RIP and RRIP for HG-GLA (p < .001). The difference between RIP/HG-GLA and RRIP/HG-GLA only approaches significance (p = 0.1178). For HG, the consequences of the parsing strategies are less substantial, but EIP does lead to small but significant gains. Finally, the pair-wise differences between OT-GLA and HG-GLA for each of the three parsing strategies are all highly significant (p < .001). This shows that with this more lenient evaluation metric, HG still has an advantage over OT in terms of end-state performance for the RIP and RRIP parsing strategies. This also means that, at least at these parameter settings, EIP/OT-GLA significantly outperforms EIP/HG-GLA, but numerically the difference is small. Overall, the same general pattern with respect to the effects of parsing strategies and choice of framework emerges as with the more strict evaluation metric reported by Jarosz (2013a). In addition, this metric also reveals that on average words are generated correctly with high accuracy even by RIP/OT-GLA while both EIP learners' performance in the limit reaches almost 100%.

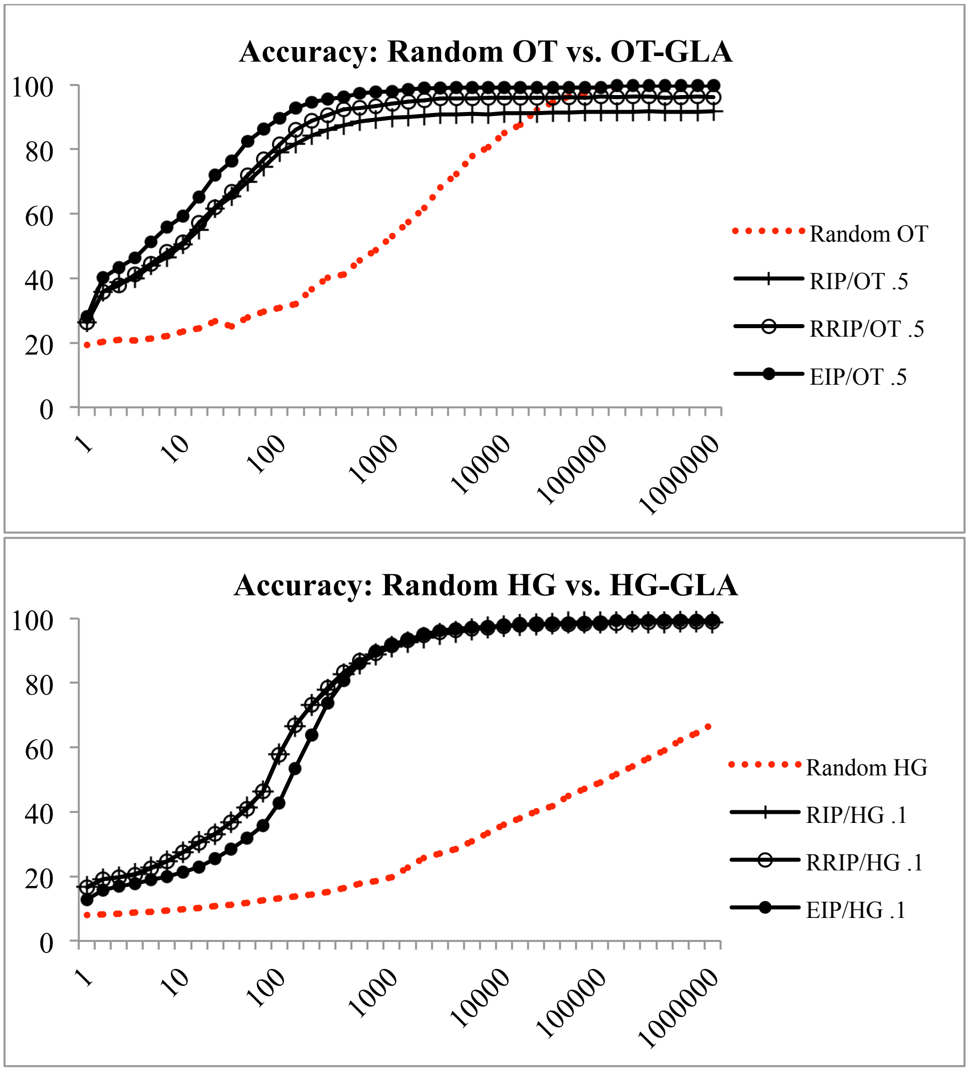
(8) Average Accuracy (SD) of GLA After 1,000,000 Iterations

Algorithm	Algorithm		
	RIP	RRIP	EIP
OT (noise 2, plasticity .5)	91.82 (0.85)	96.27 (0.67)	99.82 (0.22)
HG (noise 2, plasticity .1)	98.85 (0.11)	98.93 (0.11)	99.33 (0.11)

The learning curves for these six models are shown in Figure 3. In these figures, both the models and the baseline are evaluated using the more lenient metric. There are several observations to be drawn from these results. First, the learning curves of all the models plateau much earlier than in the previous plots. Given how high average per word accuracy is and how low per language accuracy is after about 1000 iterations, this suggests that on average the models quickly arrive at a grammar that is very close to the target grammar and, in cases where they eventually reach the target grammar, they spend a long time perfecting their grammars before they can consistently generate the stress patterns of the language. Recall that the x-axis is on a log scale: this means that, for example, EIP/OT-GLA takes about 1000 iterations to plateau in per word accuracy and then about nine times as long to plateau in per language accuracy. Qualitatively, a similar pattern holds for HG, albeit with the plateau being reached somewhat later. This suggests that the key to improving the learning curves under the stricter evaluation metric may lie in speeding up this 'perfecting' phase of learning.

Another clear result is that all the models perform much better relative to the baseline on this more lenient evaluation metric. This is good news. It means that the models are accumulating grammatical knowledge over time and getting closer and closer to the target grammars. In contrast, the random baseline fares poorly on this metric because on average it has a low chance of correctly generating a stress pattern for a word, about 20% as the initial accuracy shows. This chance of success rises very slowly because cumulative improvement requires the baseline to randomly find the target grammar (and keep it). The OT-GLA curves still rise more quickly than the HG-GLA curves, and the improved parsing strategies yield improved curves for the OT learners.

Therefore, these curves confirm the findings from the earlier sections: the parsing strategies improve not only success in the limit but the speed of learning for OT-GLA.



Accuracy: Random OT vs. OT-GLA

Accuracy: Random HG vs. HG-GLA

Figure 3 – Average Word Accuracy of OT-GLA and HG-GLA Over Time

The contribution of this section is three-fold. First, the major effects of the parsing strategies and frameworks (OT vs. HG) are consistent regardless of evaluation metric. Second, the learning curves relying on per word accuracy reveal that on average all the models quickly approach grammars that are close to the target grammars even when their final per language accuracy (requiring perfect performance on all overt forms) is fairly low. Third, the comparison between the models and the baseline on this evaluation metric gives reason to be more optimistic about their prospects to scale up. As discussed earlier, potential to scale up has to be evaluated with respect to some learning problem or task and some definition of success. This section has explored the consequences of reformulating the definition of success in a way that is perhaps more appropriate to the learning tasks that motivate stochastic models more generally. Potential to scale up depends on the definition of success, and an appropriate definition of success depends on the learning problem. The first metric considered here is probably too strict, and the second is probably overly lenient, but examining both provides a more complete understanding of the constraints on successful learning in the context of hidden structure.

No single metric can provide a comprehensive picture of the computational properties of learning models. It is therefore vital to develop our understanding of these and other learning models by evaluating them according to multiple criteria. The present investigations have considered success purely in terms of a grammar's output predictions, but further insight may also be gained by examining the structure of the acquired grammars themselves and how they differ from the target grammars. As future work considers performance of these learning models and compares OT and HG on other data sets additional questions will inevitably arise. As discussed earlier, the test set investigated here includes word types of length two to seven syllables, which provides a concrete list of words whose correctness can be evaluated according to both evaluation metrics. However, it is possible that a restriction to seven or fewer syllables is masking grammatical differences that would show up if the models and baselines were exposed to and tested on longer words. Indeed, for HG, previous work has shown that the typology may become infinite when word length is unrestricted and alignment constraints are used (Bane and Riggle to appear). Of course, all learners are only ever exposed to a finite sample of learning data, necessarily with some maximum length. This raises the difficult and general question of how to evaluate learners' acquisition of infinitely expressive grammars from finite data. How should success be defined when there are multiple distinct grammars consistent with the input data? A range of evaluation criteria must be investigated to better understand the implications of these models as well as the OT and HG frameworks for learning under various conditions.

4. Discussion

There are many considerations when evaluating the performance, feasibility and efficiency of learning algorithms. All these criteria must be considered when learning models, and the frameworks they instantiate, are compared. This paper has taken some preliminary steps toward developing a more comprehensive picture of the relative merits of stochastic OT and HG learning models facing structural ambiguity. Several directions toward a more comprehensive understanding were explored.

First, comparison with explicit random baselines is crucial. As Jarosz (2013b) showed, weak equivalence can seriously undermine the difficulty of the learning challenge posed by a test set. Neither limiting the number of iterations nor comparing the learner's processing time to the number of distinct rankings is sufficient to guarantee that an algorithm is feasible. This is because it is not known a priori how compressed a hypothesis space defined by arbitrary constraints will be made by weak equivalence, and therefore it is not known whether the amount of processing time required by an algorithm is efficient *relative to this reduced space of weakly equivalent grammars*. A random baseline provides a simple sanity check that is sensitive to the size of hypothesis space as measured by weak equivalence. The fact that this sanity check is necessary in practice, not just in principle, is evidenced by the fact that under the standard success metric explored in Section 3.2, the most widely used of the six models, RIP/OT-GLA, does not fare well. The RIP/OT-GLA curves fall consistently below the baseline, raising serious questions about the capacity of this learning model to scale to more realistic problems. It is important to emphasize, however, that the results reported here represent only a first step. Although the results look promising overall for the HG models, the HG baseline explored here may not be randomly searching the hypothesis space as effectively as possible. By sampling from a uniform grammar, weightings closer together are favored over those farther apart, and it is possible this coincidentally makes the target languages in this system harder to find for the baseline. Therefore, optimism about the HG baseline results should be reserved until the HG learners and the effects of various parameter settings can be explored more fully. More generally, future work must consider how the various learning models fare relative to a baseline when the size of the hypothesis space itself increases. That is, how well do the models compare to the baseline when the number of constraints increases? How well do the models fare on learning problems hard enough to make random search intractable? The results presented here provide reason to be optimistic since all of the models (perhaps with the exception of RIP/OT-GLA) appear to extract information from the learning data more effectively than random search in this test system. However, further systematic investigations are needed before any general conclusions can be drawn.

The second major direction explored in the present work is the comparison of learning curves of different learning models – parsing strategies in this case – to one another. These results indicate that the choice of parsing strategy has substantial consequences not only for end-state success but also for efficiency. The RRIP and EIP parsing strategies extract information from the learning data more effectively than RIP, leading to substantially faster learning for OT-GLA. This demonstrates that more principled reliance on probabilistic grammatical information has significant potential to lead to faster and more accurate models of learning. Although coping with hidden structure allows learning algorithms to grapple with more realistic learning scenarios, these simulations still massively oversimplify the true learning task facing the child. Therefore, it is of paramount importance to continue searching for more effective and more efficient methods of solving this and other learnability subproblems in order to better understand how children solve this amazingly complex task.

The third computational issue explored in this paper is the appropriate notion of success. The contrast between the results in Sections 3.2 and 3.3 highlights the significance of this choice. Under the standard, stricter notion of success, RIP/OT-GLA fares very poorly and shows little promise of scaling to more realistic problems. However, as discussed in Section 3.3, there are valid reasons to doubt the appropriateness of this strict measure of success when evaluating stochastic learning models whose goal more generally is to account not only for categorical, but also for variable, patterns and processes. The more lenient notion of success explored in Section 3.3 is not entirely appropriate either since it can obscure small but systematic divergences from the target language. However, by exploring both metrics of success, a more complete picture of the algorithms' performance emerges. The more lenient metric of success provides a much more optimistic evaluation of the models overall, with all models clearly and substantially outperforming baselines. In addition, the contrast between the algorithms' performance across the two metrics highlights a striking property of all the models explored here: the models spend relatively little time identifying grammars that are very close to the target grammars and most of their time learning how to produce a few words (or classes of words) of the target grammar. This raises the question of whether children's language acquisition has a similar character. If not, this suggests that the key to identifying further improvements in efficiency may lie in refining some of the learning strategies during this latter phase to focus the learner's efforts more effectively on the remaining cases.

References

Akers, Crystal. 2011. Simultaneous Learning of Hidden Linguistic Structures. *Simultaneous Learning of Hidden Linguistic Structures*. PhD Dissertation, Rutgers University, New Brunswick, N.J.

Alderete, John, Brasoveanu, Adrian, Merchant, Nazarré, Prince, Alan and Tesar, Bruce. 2005. Contrast Analysis Aids the Learning of Phonological Underlying Forms. In Chung-hye Han and Alexei Kochetov (editor.), *Proceedings of the 24th West Coast Conference on Formal Linguistics*, 34-42. Somerville, MA: Cascadilla Proceedings Project.

Apoussidou, Diana. 2006. On-line learning of underlying forms. *On-line learning of underlying forms*. University of Amsterdam, ms.

Apoussidou, Diana. 2007. The learnability of metrical phonology. *The learnability of metrical phonology*. PhD Dissertation, University of Amsterdam.

Apoussidou, Diana and Boersma, Paul. 2003. The learnability of Latin Stress. *IFA Proceedings* 25. University of Amsterdam. 101-148.

Bane, Max and Riggle, Jason. to appear. The typological consequences of weighted constraints. In *Proceedings of the Forty-Fifth Meeting of the Chicago Linguistic Society (2009).*

Biró, Tamás. to appear. Towards a Robuster Interpretive Parsing, Learning from overt forms in Optimality Theory. *Journal of Logic, Language and Information.*

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *IFA Proceedings* 21. University of Amsterdam. 43-58.

Boersma, Paul. 2003. Review of Tesar &amp; Smolensky (2000): Learnability in Optimality Theory. *Phonology* 20(3). 436-446.

Boersma, Paul and Hayes, Bruce. 2001. Empirical Tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32(1). 45-86.

Boersma, Paul and Levelt, Claartje. 2000. Gradual Constraint-Ranking Learning Algorithm Predicts Acquisition Order. In *Proceedings of 30th Child Language Research Forum*, 229-237. Stanford, California: CSLI.

Boersma, Paul and Pater, Joe. to appear. Convergence Properties of a Gradual Learning Algorithm for Harmonic Grammar. In John McCarthy (ed.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Coetzee, Andries and Pater, Joe. 2008a. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language &amp; Linguistic Theory* 26(2). 289-337.

Coetzee, Andries and Pater, Joe. 2008b. The place of variation in phonological theory. In John Goldsmith, Jason Riggle and Alan Yu (editor.), *The Handbook of Phonological Theory*. 2nd. Blackwell.

Daland, Robert, Hayes, Bruce, White, James, Garellek, Marc, Davis, Andrea and Norrmann, Ingrid. 2011. Explaining sonority projection effects. *Phonology* 28(02). Cambridge: Cambridge University Press. 197-234.

Dresher, Bezalel Elan. 1999. Charting the Learning Path: Cues to Parameter Setting. *Linguistic Inquiry* 30(1). 27-67.

Dresher, B Elan and Kaye, Jonathan D. 1990. A computational learning model for metrical phonology. *Cognition* 34(2). 137 - 195.

Fischer, Marcus. 2005. A Robbins-Monro type learning algorithm for an entropy maximizing version of Stochastic Optimiality Theory. *A Robbins-Monro type learning algorithm for an entropy maximizing version of Stochastic Optimiality Theory*. Master's Thesis, Humboldt University, Berlin.

Goldsmith, John. 1994. A dynamic computational theory of accent systems. In Jennifer Cole and Charles Kisseberth (editor.), *Perspectives in phonology*, 1-28. Stanford: CSLI.

Goldwater, Sharon and Johnson, Mark. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory*. Stockholm University.

Gordon, Matthew. 2002. A Factorial Typology of Quantity-Insensitive Stress. *Natural Language &amp; Linguistic Theory* 20(3). 491-552.

Hammond, Michael. 2004. Gradience, Phonotactics, and the Lexicon in English Phonology. *International Journal of English Studies* 4. 1-24.

Hayes, Bruce. 1995. *Metrical stress theory : principles and case studies*. *Metrical stress theory : principles and case studies*. Chicago: University of Chicago Press.

Hayes, Bruce. 2004. Phonological Acquisition in Optimality Theory: the Early Stages. In René Kager, Joe Pater and Wim Zonneveld (editor.), *Fixing Priorities: Constraints in Phonological Acquisition*, 245-291. Cambridge: Cambridge University Press.

Hayes, Bruce and Londe, Zsuzsa Cziraky. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23(01). 59-104.

Hayes, Bruce and Wilson, Colin. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39(3). 379-440.

Hayes, B, Zuraw, K, Siptár, P and Londe, Z. 2008. Natural and unnatural constraints in Hungarian vowel harmony. *Language.*

Heinz, Jeffrey. 2009. On the role of locality in learning stress patterns. *Phonology* 26(02). 303-351.

Jarosz, Gaja. 2006a. Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory. PhD Dissertation, the Johns Hopkins University, Baltimore, MD.

Jarosz, Gaja. 2006b. Richness of the Base and Probabilistic Unsupervised Learning in Optimality Theory. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL*, 50–59. New York City, USA: Association for Computational Linguistics.

Jarosz, Gaja. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language. Special Issue on Computational Models of Child Language Learning* 37(3). Cambridge University Press. 565-606.

Jarosz, Gaja. 2013a. Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology* 30(1). Cambridge University Press. 27-71.

Jarosz, Gaja. 2013b. Naive Parameter Learning for Optimality Theory - The Hidden Structure Problem. In Seda Kan, Claire Moore-Cantwell and Robert Staubs (editor.), *Proceedings of the 40th Annual Meeting of the North East Linguistic Society*, vol. 2, 1-14. Amherst, MA: GLSA.

Jäger, Gerhard. 2007. Maximum Entropy Models and Stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling and Chris Manning (editor.), *Architectures, rules, and preferences: variation on themes by Joan Bresnan*, 467-479. Stanford: CSLI Publications.

Jesney, Karen and Tessier, Anne-Michelle. 2011. Biases in Harmonic Grammar: the road to restrictive learning. *Natural Language and Linguistic Theory* 29(1). 251-290.

Keller, Frank. 2000. Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. PhD Dissertation, University of Edinburgh.

Legendre, Geraldine, Yoshiro Miyata and Paul Smolensky. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. CLS 26:1. 237–252.

Legendre, Géraldine, Sorace, Antonella and Smolensky, Paul. 2006. The Optimality Theory – Harmonic Grammar Connection. In *The harmonic mind : from neural computation to optimality-theoretic grammar*. Cambridge, Mass.: MIT Press.

Liberman, Mark and Prince, Allen. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8. 249-336.

Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87(4). 751-770.

McCarthy, John J and Prince, Alan. 1993. Generalized alignment. In Geert E Booij and J van Marle (editor.), *Yearbook of morphology*, 79-153. Dordrecht: Kluwer.

Merchant, Nazarré. 2008. Discovering Underlying Forms: Contrast Pairs and Ranking. *Discovering Underlying Forms: Contrast Pairs and Ranking*. PhD Dissertation, Rutgers University, New Brunswick, NJ.

Merchant, Nazarré and Tesar, Bruce. 2008. Learning underlying forms by searching restricted lexical subspaces. In *Proceedings of the Forty-First Conference of the Chicago Linguistic Society*, 33-47. Chicago Linguistics Society.

Pater, J. to appear. Canadian raising with language-specific weighted constraints. *Language.*

Pater, Joe. 2009. Weighted Constraints in Generative Linguistics. *Cognitive Science* 33. 999-1035.

Potts, Christopher, Pater, Joe, Jesney, Karen, Bhatt, Rajesh and Becker, Michael. 2010. Harmonic Grammar with Linear Programming: From linear systems to linguistic typology. *Phonology* 27(1). 1-41.

Prince, Allen. 1990. Quantitative Consequences of Rhythmic Organization. (Editor.) K Deaton, M Noske and M Ziolkowski. *CLS26-II: Papers from the Parasession on the Syllable in Phonetics and Phonology.*

Prince, Alan and Smolensky, Paul. 2004. (Optimality Theory : Constraint Interaction in Generative Grammar). Malden, MA: Blackwell Pub.

Rosenblatt, Frank. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65. 386-408.

Smolensky, Paul. 1996. The Initial State and 'Richness of the Base'. *The Initial State and 'Richness of the Base'*. Johns Hopkins University, Baltimore, MD.: Technical Report JHU-CogSci-96-4, ms.

Smolensky, Paul and Legendre, Géraldine. 2006. (The Harmonic Mind : From Neural Computation to Optimality-theoretic Grammar). Cambridge, Mass.: MIT Press.

Tesar, Bruce. 1995. Computational Optimality Theory. *Computational Optimality Theory*. PhD Dissertation, University of Colorado, Boulder, CO.

Tesar, Bruce. 1997. Multi-Recursive Constraint Demotion. *Multi-Recursive Constraint Demotion*. Rutgers University, New Brunswick, NJ, ms.

Tesar, B. 2000. Using Inconsistency Detection to Overcome Structural Ambiguity in Language Learning. *Using Inconsistency Detection to Overcome Structural Ambiguity in Language Learning*. Technical Report RuCCS-TR-58, Rutgers Center for Cognitive Science, Rutgers University., ms.

Tesar, Bruce. 2004a. Contrast Analysis in Phonological Learning. *Contrast Analysis in Phonological Learning*. Rutgers University, NJ, ms.

Tesar, Bruce. 2004b. Using Inconsistency Detection to Overcome Structural Ambiguity. *Linguistic Inquiry* 35(2). 219-253.

Tesar, Bruce. 2006a. Faithful Contrastive Features in Learning. *Cognitive Science* 30(5). 863 - 903.

Tesar, Bruce. 2006b. Learning from Paradigmatic Information. In *Proceedings of the Thirty-Sixth Conference of the North East Linguistics Society*, 619-638.

Tesar, Bruce. 2008. Output-Driven Maps. *Output-Driven Maps*. Rutgers University, New Brunswick, NJ, ms.

Tesar, Bruce. 2009. Learning Phonological Grammars for Output-Driven Maps. In *Proceedings of the Thirty-Ninth Conference of the North East Linguistics Society.*

*T*esar, Bruce, Alderete, John, Horwood, Graham, Merchant, Nazarré, Nishitani, Koichi and Prince, Alan. 2003. Surgery in Language Learning. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, 477-490.

Tesar, Bruce and Smolensky, Paul. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29(2). 229-268.

Tesar, Bruce and Smolensky, Paul. 2000. (Learnability in Optimality Theory). Cambridge, Massachusetts: MIT Press.

Tessier, Anne-Michelle. 2009. Frequency of violation and constraint-based phonological learning. *Lingua* 119(1). 6-38.

Wexler, Kenneth and Culicover, Peter. 1980. (Formal Principles of Language Acquisition). Cambridge, MA: MIT Press.

Wilson, Colin. 2006. Learning Phonology With Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science* 30(5). 945-982.