Phonological Acquisition as Weighted Constraint Interaction*

Joe Pater[1], Karen Jesney[1], and Anne-Michelle Tessier[2]
[1]University of Massachusetts, Amherst and [2]University of Alberta

1. Introduction

In the study of acquisition and learnability in Optimality Theory (OT; Prince and Smolensky 1993/2004) learning is characterized in terms of changes in constraint ranking. Learners begin with a ranking of Markedness constraints above Faithfulness constraints, and rerank them on the basis of evidence from the target language. A theory of learnability that accounts for the human acquisition process should both converge on the correct final grammar, and model the path that learners take to get there. The Gradual Learning Algorithm (GLA; Boersma 1998, Boersma and Hayes 2001) can model some, but not all aspects of the learning path, and as we will show, is non-convergent. The Constraint Demotion Algorithm (Tesar and Smolensky 1998) is convergent, but non-gradual. In this paper we argue that the search for a gradual convergent learner may be aided by replacing Optimality Theory's constraint ranking with numerical weighting, returning in this respect to OT's predecessor Harmonic Grammar (HG; Legendre *et al.* 1990, Smolensky and Legendre 2006). We demonstrate the advantages of weighting by using a minimally modified version of the GLA implemented by Boersma and Weenink (2006) that learns Harmonic Grammars, rather than OT grammars. In the next section, we provide examples of two types of gradualness in the first language acquisition of phonology, before moving on to the question of how this gradualness can be modeled in learnability theory.

2. Phonological acquisition as constraint ranking

2.1 The initial state

In OT, there are two broad families of constraints. Markedness constraints impose restrictions on Output structures (=surface phonological representations). Their content and form is generally similar to the constraints posited in other phonological theories. Some that we will be using, adopted from Prince and Smolensky (1993/2004), appear in (1):

(1) *Markedness constraints*

NOCODA	'Syllables end in a vowel'	*CVC
ONSET	'Syllables begin with a consonant'	*V
*COMPLEX	'No consonant clusters'	*CCV

Faithfulness constraints regulate the relationship between the Input (=underlying or lexical representation) and the Output form of a representation. The ranking of these OT-specific constraints with respect to Markedness determines the extent to which lexical information is retained in the face of Output restrictions. When Markedness is dominant, the ranking of Faithfulness constraints with respect to each other determines which aspect of lexical information is sacrificed. In all of our examples (as in much of child phonology), deletion brings forms into conformity with Markedness, violating MAX, defined in (2). We assume all other relevant faithfulness constraints are ranked above MAX.

* **Acknowledgments:** Thanks to Adam Albright, Diana Apoussidou, Paul Boersma, Della Chambless, Isabelle Darcy, Heather Goad, Bruce Hayes, René Kager, John McCarthy, Pat Pratt and Paul Smolensky, and members of audiences at GALANA, Utrecht University, and Potsdam University for helpful discussion.

To appear in *Proceedings of the Conference on Generative Approaches to Language Acquisition – North America (GALANA 2)*, ed. Alyona Beliokova, Luisa Meroni, and Mari Umeda. Somerville, MA: Cascadilla Proceedings Project.

(2) *Faithfulness constraint*
MAX (McCarthy and Prince 1999)
'Every element of the input must be present in the output' */CVC/ → [CV]

Children's early productions usually reduce the adult form in various ways. This can be captured with a general ranking of Markedness constraints over Faithfulness, assumed in most OT acquisition and learnability research. The tableau in (3) shows how such a ranking chooses the unmarked form, which violates only the lowest ranked constraint, over more faithful competing candidate Outputs.

(3) *Markedness >> Faithfulness in early child phonology*

/brʌʃ/ 'brush'	NOCODA	*COMPLEX	MAX
☞ [bʌ]			* *
[brʌʃ]	*	*	
[bʌʃ]	*		*

The final state in English reverses the rankings between Max and the two Markedness constraints:

(4) *Faithfulness >> Markedness in adult English*

/brʌʃ/ 'brush'	MAX	NOCODA	*COMPLEX
[bʌ]	* *		
☞ [brʌʃ]		*	*
[bʌʃ]	*	*	

When a target language has a set of Markedness constraints ranked beneath Faithfulness, the acquisition process is generally more gradual than the all-at-once change shown here. We now examine two kinds of intermediate stage displaying such gradualness.

2.2. Demotion of a subset of markedness constraints

Obviously, not all Markedness constraints are overcome at once. The order of acquisition of marked structures has been studied from Jakobson (1941/1968) through to recent research in parametric and OT frameworks. In the OT formalization of this type of intermediate stage, the Markedness constraints violated by acquired structures are ranked beneath conflicting Faithfulness constraints, while other Markedness constaints continue to dominate at least one Faithfulness constraint. An example of this configuration appears in (5), in which NOCODA but not *COMPLEX has been demoted beneath MAX, thus allowing codas, but not clusters to surface.

(5) *Intermediate stage 1: M1 >> F >> M2*

/brʌʃ/ 'brush'	*COMPLEX	MAX	NOCODA
☞ [bʌʃ]		*	*
[bʌ]		* *	
[brʌʃ]	*		*

Trevor (Pater 1997) in fact produced the intermediate form (6b) for a period of nearly six months.

(6) *Gradual acquisition of syllable structure*
a. [bʌ] 'brush' 1;5.18 b. [bʌʃ] 'brush' 1;6.17 – 2;0.3 c. [bɝʌʃ] 'brush' 2;4.13

These limited data are representative of a stage that most, if not all, Dutch and English-learning children go through. In a study of longitudinal data from 12 Dutch children, Levelt and van de Vijver

(2004) found that they all demoted NoCoda before *Complex. They further show that the order of demotion correlates with frequency of syllable type in a corpus of child-directed speech containing 112,926 primary stressed syllables:

(7) *Percentage of syllables violating constraints in Dutch child-directed speech*

NoCoda	49.95 %
*Complex-Onset	3.62 %

While this correlation between frequency and order of acquisition does not in itself demonstrate causation, several studies show differences in acquisition order between languages in which the relative frequencies of the structures also differ (e.g. Ingram 1988, Roark and Demuth 2000). Such cross-linguistic evidence strongly argues that frequency can affect order.

2.3 Partial demotion of a constraint

In OT, a constraint can occupy a status between full satisfaction in the initial state, and inactivity in the final state (Barlow 1997, Pater 1997, Gnandesikan 2004). In such an intermediate stage, the effects of the constraint are seen in some phonological environments, but not others. One sub-type of this phenomenon occurs when a Markedness constraint is demoted beneath a Positional Faithfulness constraint (Beckman 1998) but not its general counterpart. A number of recent studies have provided acquisition data corresponding to this sort of ranking (see Tessier 2006 for an overview, and a defense of the Positional Faithfulness analysis against alternatives). One example comes from Rose's (2000) study of the acquisition of Québecois French, in which Clara passes through a stage that allows clusters in stressed but not unstressed syllables. The examples in (8a.) include faithfully produced clusters in the final syllable, which is stressed in French, while (8b.) exemplifies deletion in the initial syllable of disyllables, which is unstressed.

(8)	*Child form*	*Adult form*	*Word*	*Gloss*	*Age*
a.	[klɪs]	[glɪs]	*glisse*	's/he slides'	1;10.4
	[paˈpχɔ]	[bibʁɔ̃]	*biberon*	'baby bottle'	1;9.29
b.	[biˈle]	[bʁyle]	*brulé*	'burned'	1;9.29
	[ka'sæd]	[glisad]	*glissade*	'a slide'	1:10.4

Rose's analysis of these data invokes a specific version of the Max constraint (Goad and Rose 2004):

(9) Max-(Stress) Every element in an input stressed syllable must be present in the output

The tableaux in (10) show that a ranking of this constraint above the Markedness constraint, which in turn continues to dominate the broader Max constraint, produces the pattern in question.

(10) *Max-(Stress) >> *Complex >> Max*

/glɪs/ 'glisse'	Max-(Stress)	*Complex	Max
☞ [klɪs]		*	
[kɪs]	*		*

/glisad/ 'glissade'	Max-(Stress)	*Complex	Max
☞ [kasæd]			*
[klasæd]		*	

Again, these data from one child are reflective of a broader tendency. Chambless (2006) conducted a nonce-word production study with 18 English learning children aged 2;3 to 3;3 (mean 2;8) in which clusters were included in word-initial stressed position, word-medial stressed position, and word-initial unstressed position. Both word-initial and stressed positions are common environments for Positional

Faithfulness constraints. Children produced cluster targets as clusters more often in initial stressed position (74%) than in either initial unstressed (37%) or medial stressed (62%) positions. A repeated measures ANOVA found a main effect for position, $F(2,26)= 21.19$, $p < .001$. Planned contrasts found significant differences between initial-stressed and initial-unstressed positions, $F(1,13) = 34.47$, $p < .001$, as well as initial-stressed and medial-stressed conditions, $F(1,13) = 7.950$, $p=.014$. While these are cross-sectional rather than longitudinal data, they strongly suggest that children do often acquire clusters first in the phonological environments to which Positional Faithfulness applies.

3. The Gradual Learning Algorithm

Part of the appeal of OT is that it comes equipped with a convergent learning algorithm, the Constraint Demotion Algorithm (CDA; Tesar and Smolensky 1998). It is convergent in the sense that if a language can be generated by one or more rankings of a set of constraints, the CDA is guaranteed to find one of those rankings. This holds when the learner is provided with all necessary information about 'hidden structure', including the correct underlying representations, and the correct morphological and prosodic parse of the surface strings. Learning of hidden structure is a topic of much current research - see Tesar (1998) *et seq.* for proposals that make use of the ability of the CDA to detect inconsistency, and Jarosz (2006) and Apoussidou (2007) for alternatives.

However, the CDA in its standard form does not learn gradually in any way that could model the acquisition paths taken by human learners (cf. Tessier 2006). This problem forms part of the motivation for Boersma's (1998) alternative proposal, the GLA. The GLA can learn gradually because it represents the current state of the grammar numerically, and learning consists of gradual adjustment of these numerical values. Boersma and Levelt (2000) make use of this aspect of the GLA in modeling Levelt and van de Vijver's (2004) Dutch syllable structure data. In the following, we report a replication of part of that result using the GLA implementation in Praat (Boersma and Weenink 2006).

3.1 Demotion of a subset of Markedness constraints in the GLA

The first step in the GLA, as in the error-driven version of the CDA, is to parse an incoming piece of data using the current state of the grammar. The learner is supplied with an Input-Output pair (in the cases under consideration, the Input is simply identical to the observable Output), and it finds the Output that its own grammar produces for that Input. If the resulting Output does not match the observed Output, learning is triggered.

As in Boersma and Levelt (2000), we assume an initial ranking of Markedness constraints at 100, and Faithfulness at 50:

(11) *Initial State*

*COMPLEX	100
NOCODA	100
MAX	50

The numerical values of the GLA's constraints are converted to an OT ranking each time the grammar is used. The values in (11) would be converted to the ranking in (12a). Given a learning datum with a cluster and a coda, like *praat* 'speak', this grammar would produce a cluster-less, coda-less Output, as in (12c) (see also the equivalent tableau in (3)).

(12) *Parsing with initial state grammar*

a. Ranking: *COMPLEX, NOCODA >> MAX
b. Learning datum: /pra:t/ → [pra:t]
c. Parsing: /pra:t/ → [pa:]

Since the resulting Output [pa:] does not match the observed Output [pra:t], learning is triggered. In both the CDA and the GLA, learning consists of adjusting the grammar in the direction of prefering the observed form (the 'Winner' as Prince 2002a terms it) over the learner's current Output (the 'Loser').

The table in (13) shows the violation marks incurred by each member of the Winner (W) - Loser (L) pair.

(13) *Winner/Loser pair*

/pra:t/	NoCoda	*Complex	Max
L [pa:]			* *
W [pra:t]	*	*	

The CDA adjusts the ranking by demoting all of the constraints assessing a violation mark to the Winner beneath some constraint assessing a violation mark to the Loser. This results in the grammar correctly preferring the Winner, but does not allow for gradual learning. The GLA, on the other hand, adjusts the grammar as in (14) (this is the 'symmetric all' default setting in Praat, which follows the formulation used by Boersma and Levelt 2000, Boersma and Hayes 2001).

(14) Demote all constraints assessing a violation mark to the Winner by x, promote all constraints assessing a violation mark to the Loser by x

The value of x, termed plasticity, is adjustable in Praat. We adopted the value of 0.1 from Boersma and Levelt (2000), which when inserted into (14) results in the following grammar:

(15)	*Complex	99.9	/pra:t/	→ [pa:]
	NoCoda	99.9		
	Max	50.1		

The learner will continue to make parsing errors until sufficient data accumulate to push the Markedness and Faithfulness constraints into the reverse order.

The learner is therefore gradual; it is also sensitive to frequency, since the speed at which a Markedness constraint's value changes depends on how often violations are encountered in learning trials. After 800 trials with data distributed as in (7), and standard Praat GLA settings except for plasticity, the grammar is at the intermediate stage where it parses codas but not clusters faithfully:

(16)	*Complex	95.89	/pra:t/ → [pa:t]
	Max	84.92	
	NoCoda	75.21	

After 2400 trials, it parses both structures faithfully, as in the adult language:

(17)	Max	94.43	/pra:t/ → [pra:t]
	*Complex	92.83	
	NoCoda	75.21	

3.2 Positional Faithfulness and the GLA

Thus, the GLA can model the demotion of a subset of Markedness constaints, and the connection between frequency and order of acquisition. However, the GLA will not pass through the second kind of intermediate stage discussed in Section 2, where a Markedness constraint has been demoted beneath a Positional Faithfulness constraint, but not its general counterpart. As discussed in Tessier (2006), this is due to the fact that violations of the specific Faithfulness constraint are a subset of those of the general one. Taking the Max-Stress and Max example, Max is violated by all cases of deletion, while Max-Stress is violated only in deletion in stressed syllables Because of this subset relation, the general Faithfulness constraint (e.g. Max) will always be promoted more quickly than the specific one (e.g. Max-Stress), regardless of the frequency of violation in the two environments.

To illustrate this, we present the results of another simulation using the same GLA settings as for the previous one, but with the two MAX constraints, and *COMPLEX. We provided the learner with an equal distribution of words with stressed clusters (i.e. *glisse*) and unstressed clusters (*glissade*). At 200 trials, the general MAX constraint is well ahead of the Positional Faithfulness constraint. It will eventually overcome *COMPLEX first, and the grammar will parse all clusters correctly, without having passed through the intermediate stage shown in (10).

(18) *No intermediate specific Faithfulness stage*

After 200 trials:	*COMPLEX	80.1	/glis/ → [kis]
	MAX	69.9	/glisæd/ → [kasæd]
	MAX-STRESS	59.8	

The specific-general faithfulness relation is also a problem for correctly reaching the final state, when learning a language that only permits the structure in the specific context, for both the CDA and the GLA (Smith 1999, Hayes 2004, Prince and Tesar 2004, Tessier 2006). The following shows the end-state for the acquisition of a language with clusters only in stressed syllables (as in southeastern Brazilian Portuguese; Harris 1997, Goad and Rose 2004), using Praat's default GLA settings, with an initial value of 100 for the markedness constraint, and 50 for the faithfulness constraints. The learning data consisted of an equal distribution of the two mappings shown alongside the ranking values.

(19) *Superset grammar*

Final State:	MAX	80.3	/'CCV/ → ['CCV]
	MAX-STRESS	80.3	/CV.'CCV/ → [CV.'CCV]
	*COMPLEX	69.7	

This grammar does correctly parse the forms of the language, but because MAX is ranked over *COMPLEX, it also parses ungrammatical forms, as shown in (20) in a 'Richness of the Base' tableau (Prince and Smolensky 1993/2004), which provides an input with a cluster in an unstressed syllable:

(20) *Restrictiveness problem*

CCV.'CV	MAX-(STRESS)	MAX	*COMPLEX
☞ CCV.'CV			*
CV.'CV		*	

A properly restrictive learner would reach a final state in which the grammar permits clusters only in the environment in which they are present in the learning data, that is, in stressed syllables.

One might question the assumptions underlying the present example, which was chosen primarily because it so closely mirrors the intermediate acquisition stage. A non-standard aspect of the analysis is that it requires the presence of underlying stress in a language with predictable stress. One could instead invoke a Markedness constraint specific to unstressed syllables, which would not only be compatible with a lack of underlying stress, but would also avoid the restrictiveness problem. However, the restrictiveness problem is a very general one that arises in all instances where Positional Faithfulness is used to capture an asymmetry in a language's inventory. Prince and Tesar's (2004) response to it is in fact to suggest that Positional Faithfulness be abandoned, but it does not appear that Positional Faithfulness can always be replaced by specific Markedness constraints (see Tessier 2006 for recent discussion; see also Boersma 2006 for a very different way of eliminating Positional Faithfulness).

Besides banning Positional Faithfulness constraints from the theory, another possibility is for the learner to calculate the specific-general relation when deciding which constraint to install in a stratum (Hayes 2004, Tessier 2006). However, this solution is cast in a modified version of the CDA, and it is hard to see how it could be incorporated into the GLA. A further possibility is to adopt a learning algorithm that provides the learner with the distribution of observed forms, along with the set of possible underlying forms (or a means of generating them), and which maximizes the probability of the observed forms (Jarosz 2006). However, it remains to be seen if such an approach can scale up to realistic learning scenarios. In section 4, we will provide an alternative solution that simply alters the parsing mode used by the GLA. Before doing so, we discuss some further issues with the GLA that our proposal resolves.

3.3 Non-convergence in the GLA

Dresher (1999) draws attention to the Credit Problem in learning, which arises in any theory with principle/rules/constraints of any generality: Given multiple solutions for the problem at hand, how does the learner know which one to choose?

An OT version of the Credit Problem is schematized in (21). In this discussion we will adopt Prince's (2002a) notation, in which each row of a tableau contains a Winner-Loser pair, and constraints are tagged for whether they prefer the Winner or Loser. A corresponding standard tableau is also provided for comparison.

(21) *OT credit problem schematized*

	W ~ L	Con1	Con2	Con3
Input	$\text{Output}_W \sim \text{Output}_L$	W	L	W

Input	Con1	Con2	Con3
Winner		*	
Loser	*		*

The correct ranking must place Con1 or Con3 over Con2 – but which one? A decision can be made based on a further piece of data, added to the original one in (22).

(22) *Credit problem resolved* (1 >> 2 >> 3)

	W ~ L	Con1	Con2	Con3
In-1	$\text{Out-1}_W \sim \text{Out-1}_L$	W	L	W
In-2	$\text{Out-2}_W \sim \text{Out-2}_L$		W	L

Because the second form requires the ranking Con2 >> Con3, the first form must be dealt with in terms of Con1 >> Con2. The CDA is guaranteed to find this solution, given sufficient data. However, unlike the CDA, the GLA does not use ranking logic, and will not necessarily realize how to resolve such a problem. In fact, by iterating this pattern just a few times, we can create a scenario that will consistently foil the standard 'symmetric all' version of the GLA. This scenario, which involves four pieces of learning data, is schematized in (23).

(23) *An iterated resolved credit problem* (1 >> 2 >> 3 >> 4 >> 5)

	W ~ L	Con1	Con2	Con3	Con4	Con5
In-1	$\text{Out-1}_W \sim \text{Out-1}_L$	W	L	W		
In-2	$\text{Out-2}_W \sim \text{Out-2}_L$		W	L	W	
In-3	$\text{Out-3}_W \sim \text{Out-3}_L$			W	L	W
In-4	$\text{Out-4}_W \sim \text{Out-4}_L$				W	L

A typical outcome, produced with equal distribution of the 4 Input/Winner mappings, constraint values starting at 100, and default Praat GLA settings, is shown in (24).

(24)

Constraint	*Ranking Value*
Con3	12392.505
Con4	12392.208
Con1	12391.368
Con2	12391.023
Con5	12390.103

Values this high are indicative of non-convergence - the ranking values are cycling higher and higher, without settling on a set of values that will produce the observed data. If the GLA is run again, the values will simply continue to rise. The values are also still very close to one another. One aspect of the GLA that we have yet to discuss is that the value that is used to map to a ranking is not exactly the ranking value, but instead a random sampling from a normal distribution around the ranking value. This is used to model variation. When two ranking values are close to one another and their corresponding normal distributions overlap considerably, their ranking will vary in repeated samplings. In the present case, this produces variation for all of the Inputs. A test run giving this grammar each Input 100000 times, with the evaluation 'noise' setting of 2 (that is, the standard Praat settings for 'To Output distributions'), produces the percentage of correct forms seen in (25):

(25)	Input1 → Winner	79.4%
	Input2 → Winner	55.8%
	Input3 → Winner	57.6%
	Input4 → Winner	77.8%

One may wonder if the data pattern in (23) occurs in natural languages. It is clear that the WLW pattern does exist for particular sets of constraints and candidates, since the pair of Ws will be produced anytime two constraints with the same preference overlap in their domain of evaluation. Whether just this pattern turns up in learning data is in fact difficult to know, which is in itself a problem. Without a precise characterization of the types of constraints that will give rise to this pattern, one could not ban them, as in Prince and Tesar's (2004) response to the Positional Faithfulness problem. One could also not make predictions about which sorts of languages should be unlearnable (see Boersma 2003 for relevant discussion).

There is some indication, however, that this problem crops up in a less severe form in examples modeled on natural language. The abstract case in (23) was created to explore a hypothesis about the source of the difficulties the GLA has in learning a portion of the Hungarian vowel harmony pattern discussed in Hayes and Londe (2006). The generalizations present in the data were as follows:

(26) *Fragment of Hungarian vowel harmony* (based on Hayes and Londe 2006)

a. Backness harmony applies to suffixes, but not roots: suffixes assimilate obligatorily to a root-final back or front rounded vowel

b. A root that ends in a neutral (front unrounded) vowel preceded by a back vowel permits both back and front suffixes

c. A root that ends in a neutral vowel preceded by a front rounded vowel allows only a front suffix

In a departure from Hayes and Londe (2006), we assume the alternating suffix vowel is underlyingly [+back], and gave that information to the learner. The constraint set was as in (27)

(27) *Constraints* (slightly adapted from Hayes and Londe 2006)

AGREE-B	Adjacent vowels agree in backness if one is round
IDENT-B	A segments' backness value is identical in Input and Output
IDENT-B-ROOT	A root segments' backness value is identical in Input and Output
AGREE-DIST-B	A back vowel is followed by only back vowels
AGREE-LOC-N	A neutral vowel is followed by an adjacent front vowel
AGREE-DIST-F	A front rounded vowel is followed by only front vowels

There is considerable overlap in the scope of the constraints. For example, when an underlying [+back] suffix follows a sequence of front vowel followed by a neutral vowel, both AGREE-LOC-N and AGREE-DIST-F prefer the correct surface [-back] vowel. In fact, the constraint set is richer than needed for the generalizations in (26), but this is hardly an issue, since a universal constraint set will always contain some redundancy.

By applying ranking logic, we can arrive at a set of necessary dominance relationships between the constraints. These are shown numerically in the first column of values in (28). The values were also chosen to be far enough apart to avoid variation. The second set of values show the result of applying the GLA with the standard Praat settings (see Pater to appear for the dataset and further details).

(28) *Ranking values arrived at through ranking logic and by applying the GLA*

IDENT-B-ROOT	100	1093.152
AGREE-B	85	1088.086
AGREE-DIST-F	70	1084.670
IDENT-B	55	1082.421
AGREE-N	40	1079.287
AGREE-DIST-B	40	1064.816

If we apply the 'get fraction correct' function of the GLA, the handcrafted values usually produce 100% correct over the whole set of inputs. The GLA-produced values, however, yield percentages correct as low as 79.3 for some of the inputs. The relative order matches the handcrafted values, but they are too close together, thus producing unwanted variation. The high values again indicate that the algorithm is not converging, and iterated runs result in continued increase. This outcome is typical for this problem: 10 trials using standard settings in Praat for learning and 'get fraction correct' produced an average 93.4% correct with the best being 95.3%.

4. Acquisition and learnability with weighted constraints

The GLA derives several benefits for the modeling of learning from its incorporation of numerical values for constraints. We have focused on (29a.); see further Boersma and Levelt (2000) and many others on this point and (29b.), and see especially Boersma and Hayes (2001) on (29c.).

(29) a. Learning can be gradual, and affected by frequency
b. Variation is predicted to occur between 'stages'
c. Variation in the target language can be learned

In this section we show that further benefits may accrue from adopting a pure numerical version of the GLA, in which no conversion to ranking is made, and candidates are evaluated by using numerically weighted constraints. Both the learning path and final state convergence problems raised for the GLA in the section 3 can be resolved in this way.

The version of weighted constraint evaluation we adopt is one suggested as a strawman in Prince and Smolensky (1993/2004), and which has been further explored in Prince (2002b), Keller (2006), and Pater *et al.* (2006):

(30) a. Each violation is multiplied by the weight (=ranking value) of the constraint
b. A candidates' score is the sum of the weighted violations
c. The optimal candidate has the lowest score

Assessing the result of replacing ranking with weighted evaluation is greatly facilitated by the fact that Praat offers this evaluation mode as an option, which it terms 'LinearOT', after Keller (2006). We will instead use the abbreviation HG-GLA to refer to the GLA with this modification, to emphasize the absence of ranking in this model, and the closeness to Harmonic Grammar.

4.1 Gradualness and restrictiveness in the HG-GLA

In section 3.2, we discussed two related problems for the GLA: it does not predict an intermediate stage in which only the positional version of a Faithfulness constraint outranks Markedness, and it does not correctly acquire final state grammars with this ranking configuration.

To show the effect of a switch to weighting, we ran the same simulation for the learning of onset clusters as described above, with only the evaluation mode changed to <LinearOT>. After 250 trials, the constraint values were as indicated in the 'Weights' row of the weighted constraint tableau in (31). The final column indicates the scores of the candidates calculated as in (30), and the optimal candidate is the one with the lowest such score. Here the result is the same as it would be for a ranked tableau.

(31) *No clusters in unstressed syllables*

Weights	88.3	61.7	50.4	Σ
/glisæd/ 'glissade'	*COMPLEX	MAX	MAX-STRESS	
☞ [kasæd]		*		61.7
[klasæd]	*			88.3

However, in a stressed syllable, both MAX and MAX-STRESS apply. In this case we see a different outcome in the weighted constraint system: the simultaneous violation of both lower-valued constraints is worse than a single violation of the higher-valued constraint.

(32) *Clusters in stressed syllables*

Weights	88.3	61.7	50.4	Σ
/glis/ 'glisse'	*COMPLEX	MAX	MAX-STRESS	
☞ [klis]	*			88.3
[kis]		*	*	112.1

The possibility of this sort of 'gang effect', where two or more lower-valued constraint violations overcome a higher one, is what distinguishes weighted from ranked constraint systems. Here the gang effect yields a positive result, in that it allows the GLA to pass through this sort of intermediate stage.

If the learning data consists only of clusters in stressed syllables, the learner will stop making errors once the summed value of the two faithfulness constraints overcomes *COMPLEX. If the previous simulation is run with this change, and with standard Praat settings, the final state is as in (33).

(33) *Final state - clusters in stressed syllables only*

*COMPLEX	97.9
MAX	52.1
MAX-STRESS	52.1

These values produce the same results if inserted into the tableaux in (31) and (32). The tableau in (32) would represent an attested form of the language. The tableau in (31) would be a Richness of the Base tableau showing that a cluster in an unstressed syllable is reduced, as desired of a restrictive system.

While this result is encouraging, it is important to note that using weighting in the GLA is not a panacea for the restrictiveness problems it faces (see Hayes and Londe 2006 and Tessier 2006 for others). In particular, gang effects between Faithfulness and Markedness can still produce superset grammars. For further discussion of these issues, see Jesney, Pater and Tessier (in prep.).

4.2 Convergence in the HG-GLA

To assess the effect of a switch to weighting on the convergence properties of the GLA, we presented the same learning problems as in 3.3 to the Praat implementation of the GLA with default settings, except for the evaluation mode being set to <LinearOT>. For the WLW learning problem in (23), the average percentage correct obtained using 'get fraction correct' after each of 10 learning trials was 100%. This is not surprising, since there is very little learning required to produce this pattern under weighted evaluation; the initial weighting will correctly parse 3 of the 4 mappings, as shown in (34).

(34) *Parsing under initial weighting*

Weights	100	100	100	Σ
Input	Con1	Con2	Con3	
☞ Winner		*		100
Loser	*		*	200

To give the HG-GLA a more parallel test, we created the system illustrated in (35), in which there is the same sort of iterated conflict between mark-data pairs, but which the initial state will only parse at about 50% correct.

(35) *Weighting-specific learning problem*

	W ~ L	Con1	Con2	Con3	Con4	Con5	Con6
In-1	Out-1_W ~ Out-1_L	W	L	W	L		
In-2	Out-2_W ~ Out-2_L		W	L	W	L	
In-3	Out-3_W ~ Out-3_L			W	L	W	L
In-4	Out-4_W ~ Out-4_L				W	L	W

Out of ten trials with the same conditions, the lowest overall percentage correct obtained was 99.995%.

For "Hungarian", the lowest overall percentage correct obtained out of ten trials was 99.997%.

There are two likely reasons why a version of the GLA with weighted evaluation may have higher convergence rates. First, the conversion from numerical values to rankings in the standard GLA introduces a non-linearity in the evaluation/parsing system that is not reflected in the learning system. Put another way, a successful learning algorithm for OT grammars likely needs to reflect the mode of constraint interaction, as in the CDA. Second, the credit problem takes on a different form in weighted systems, since credit can be distributed amongst constraints, as in the gang effects shown here.

5. Conclusions

The GLA in its standard formulation takes a large step forward in meeting the goal of having a gradual and convergent learner, including one that works in noise. Here we have suggested that replacement of OT evaluation with HG weighted evaluation may result in further progress in modeling both gradualness and convergence. Future research should include more systematic testing of both the OT-GLA and the HG-GLA under a range of conditions, along with a comparison of these models with other weighted constraint learners, especially the Maximum Entropy models of Goldwater and Johnson (2003), Jäger (to appear) and Wilson (to appear). As these authors emphasize, one of the primary advantages of adopting linguistic grammars with weighted constraint interaction is that the study of learnability can draw on a wealth of research into the learning of such systems (see also Pater *et al.* 2006 for a weighted constraint learner that resembles in some ways the CDA, and for arguments that weighted interaction does not result in an overly powerful theory of grammar).

Weighted constraint systems also open up new areas of research in the study of acquisition. As we have shown, gang effects emerge as a natural consequence of gradual learning, even when such gang effects are not present in the target language. This leads to the predicted intermediate stages beyond those that are yielded by systems with ranking.

References

Apoussidou, Diana. 2007. *The Learnability of Metrical Phonology*. Ph.D. dissertation, University of Amsterdam.

Barlow, Jessica. 1997. *A Constraint-Based Account of Syllable Onsets: Evidence from Developing Systems*. Ph.D. dissertation, Indiana University.

Beckman, Jill. 1998. *Positional faithfulness*. Ph.D dissertation. Amherst: University of Massachusetts.

Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Ph.D. dissertation, University of Amsterdam.

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

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

Boersma, Paul. 2006. The acquisition and evolution of faithfulness rankings. Handout from the Manchester Phonology Meeting (http://www.fon.hum.uva.nl/paul/presentations/)

Boersma, Paul and Clara C. Levelt. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. In *The proceedings of the thirtieth annual child language research forum*, ed. Eve V. Clark. Stanford: CSLI.

Boersma, Paul and David Weenink. 2006. Praat: doing phonetics by computer. Software version 4.5.1. Retrieved October 28, 2006 from www.praat.org.

Chambless, Della. 2006. Asymmetries in the acquisition of consonant clusters. Ph.D. dissertation, University of Massachusetts, Amherst.
Dresher, Bezalel Elan. 1999. Charting the Learning Path: Cues to Parameter Setting. *Linguistic Inquiry* 30: 27-67.
Gnanadesikan, Amalia. 2004. Markedness and Faithfulness Constraints in Child Phonology. In Kager *et al.* (eds.)
Goad, Heather, and Yvan Rose. 2004. Input Elaboration, Head Faithfulness and Evidence for Representation in the Acquisition of Left-Edge Clusters in West Germanic. In Kager *et al.* (eds.)
Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*.
Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the Early Stages. In Kager et al. eds.
Hayes, Bruce and Zsuzsa Cziráky Londe. 2006. Stochastic Phonological Knowledge: The Case of Hungarian Vowel Harmony. Phonology 23:59-104
Ingram, David. 1988. The Acquisition of Word-Initial [v]. Language and Speech. 31(1). 77-85.
Jäger, Gerhard. To appear. Maximum entropy models and Stochastic Optimality Theory. In Architectures, rules, and preferences: A festschrift for Joan Bresnan, ed. Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, and Annie Zaenen. Stanford, CA: CSLI.
Jakobson, Roman. 1941/1968. Child Language, Aphasia and Phonological Universals. Trans. R. Keiler. The Hague: Mouton.
Jarosz, Gaja. 2006. *Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory*. Ph.D. dissertation, Johns Hopkins University. ROA-884.
Jesney, Karen, Joe Pater, and Anne-Michelle Tessier. In prep. Restrictiveness and gang effects in learning. Ms., University of Massachusetts, Amherst, and University of Alberta.
Kager, René, Joe Pater and Wim Zonneveld. 2004. Constraints in Phonological Acquisition. Cambridge University Press.
Keller, Frank. 2006. Linear Optimality Theory as a Model of Gradience in Grammar. In *Fanselow et al.* eds., *Gradience in Grammar: Generative Perspectives*. Oxford University Press.
Legendre, Geraldine, Yoshiro Miyata, and Paul Smolensky. 1990a. Harmonic Grammar – a formal multi-level connectionist theory of linguistic wellformedness: An application. In *Proceedings of the twelfth annual conference of the Cognitive Science Society*, 884–891. Cambridge, MA: Lawrence Erlbaum.
Legendre, Geraldine, Antonella Sorace, and Paul Smolensky. 2006. The optimality theory–harmonic grammar connection. In Smolensky and Legendre (eds.)
Levelt, Claartje, and Ruben van de Vijver. 2004. Syllable types in cross-linguistic and developmental grammars. In Kager *et al*. (eds.)
McCarthy, John and Alan Prince. 1999. Faithfulness and Identity in Prosodic Morphology. In René Kager, Harry van der Hulst, and Wim Zonneveld (eds.). *The Prosody Morphology Interface*. Cambridge: CUP.
Pater, Joe. 1997. Minimal violation and phonological development. *Language Acquisition* 6. 201-253.
Pater, Joe. To appear. Non-convergence in the Gradual Learning Algorithm. *Linguistic Inquiry*.
Pater, Joe and Jessica Barlow. 2003. Constraint Conflict in Cluster Reduction. Journal of Child Language 30, 487-526.
Pater, Joe, Christopher Potts, and Rajesh Bhatt. 2006. *Harmonic Grammar with Linear Programming*. Ms, University of Massachusetts, Amherst.
Prince, Alan. 2002a. Arguing optimality. In *Papers in Optimality Theory* II, ed. Andries Coetzee, Angela Carpenter, and Paul de Lacy, 269–304. Amherst, MA: GLSA. ROA-562.
Prince, Alan. 2002b. Anything goes. In New century of phonology and phonological theory, ed. Takeru Honma, Masao Okazaki, Toshiyuki Tabata, and Shin ichi Tanaka, 66–90. Tokyo: Kaitakusha. ROA-536.
Prince, Alan, and Paul Smolensky. 1993/2004. Optimality Theory: Constraint interaction in generative grammar. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Revised version published by Blackwell, 2004.
Prince, Alan and Bruce Tesar. 2004. Learning phonotactic distributions. In Kager *et al*. eds.
Riggle, Jason. 2006. Lexical entropy, finite state optimality, and learning from surface forms alone. Colloquium, University of Massachusetts, Amherst.
Roark, B. and Katherine Demuth. 2000. Prosodic constraints and the learner's environment: A corpus study. Proceedings of the Boston University Conference on Language Development, 24, 597-608.
Smolensky, Paul, and Geraldine Legendre. 2006. *The harmonic mind: From neural computation to Optimality-Theoretic grammar*. Cambridge, MA: MIT Press.
Tesar, Bruce. 1998. Using the mutual inconsistency of structural descriptions to overcome ambiguity in language learning. In Proceedings of NELS 28, ed. P. Tamanji and K. Kusumoto, 469-483. Amherst, MA: GLSA, University of Massachusetts.
Tesar, Bruce and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29. 229-268.
Tessier, Anne-Michelle. 2006. *Biases and stages in phonological acquisition*. Ph.D. dissertation, University of Massachusetts, Amherst.
Wilson, Colin. To appear. Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization. In *Cognitive Science*