

Frequency of violation and constraint-based phonological learning

Anne-Michelle Tessier

University of Alberta, 4-22 Assiniboia Hall, Edmonton, Alberta, Canada T6G 2E7

Received 6 June 2006; received in revised form 23 June 2008; accepted 28 July 2008

Available online 20 September 2008

Abstract

This paper provides two arguments that error-driven constraint-based grammars should not be learned by directly mirroring the frequency of constraint violation and satisfaction in the target words of a language. The first argument comes from a class of stages attested in phonological development, called Intermediate Faith (IF) stages, in which children produce marked structures only in privileged positions. Two such stages are presented and analyzed, from the literature on English and French L1 acquisition, and their learning consequences are examined. The second argument concerns the degree of restrictiveness that a learner's end-state grammar encodes, using two hypothetical interactions between learner's assumptions about hidden structure and developing constraint rankings that can trick a learner into adopting a superset grammar. These two arguments are used to support an approach called Error-Selective Learning (ESL), in which errors are learned and stored gradually, in a way that relies on violation frequency, but rankings themselves are learned in a *non-gradual* way (relying on the algorithms of Prince and Tesar, 2004; Hayes, 2004). It is also shown that violation frequencies can still cause problems regardless of a learner's method of grammatical evaluation—either ranked constraints as in Optimality Theory, or weighted constraints as in Harmonic Grammar.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Phonological acquisition; Learnability theory; Optimality theory; Harmonic grammar; Developmental stages; Frequency; Constraint re-ranking; Faithfulness constraints; Subset principle

1. Introduction

Current theories of learning in constraint-based theories like Optimality Theory (Prince and Smolensky, 1993/2004), and a substantial body of work on L1 and L2 phonological development, are beginning to allow researcher to assess the match between learnability theories and empirical predictions. The recent literature has seen a large body of work whose main goal is to characterize, using OT tools, the pre-target production grammars of L1 learners (beginning with, e.g. Gnanadesikan, 1995/2004; Demuth, 1996; Pater, 1997). At the same time, OT learnability work has sought to answer a different but related question: how to build a learner that uses the available errors to reach the right target final grammar (see references throughout this paper). In nearly all of this latter work, the learner is assumed to be error-driven, meaning that it proceeds from each stage to the next by making errors, comparing its errors to the target forms, and changing its grammar in some way as a result.

This paper is part of a research program that seeks to integrate results from both domains, by asking to what extent learners that are successful from a learnability perspective are also good at replicating the human learning process. How do real learners get from one time-slice grammar to the next? Some learnability work is very careful to not make

E-mail address: amtessier@ualberta.ca.

claims about how humans learn, and fairly so—but Optimality Theory (and other constraint-based grammars) are proposed, in the generative tradition, as a way of capturing what humans know about their native language phonology, and learnability insights about constraint-based grammars are also insights into the properties of a system that human learners must somehow come to know.

The particular issue that this paper addresses is what makes a constraint-based learner gradual: that is, what mechanism in the learning process ensures that learners move incrementally from the initial to final states, and how the stages in between are determined. I will consider two potential answers. The first, which might be viewed as the default assumption, is that learners are gradual because they re-rank their constraints in a gradual fashion. In this approach, constraints must be given numerical values of some kind; on the basis of each error made by the learner, these values are slowly brought closer together or farther apart and over time approximate the end-state grammar. This approach is fairly simple, frequently successful, and appealing in its straightforward account of various aspects of learning, notably variation—however, I will argue against it. Instead, I will suggest that the gradualness of grammar learning should come from the gradual incorporation of errors into the learner’s system, and that the method of re-ranking on the basis of any particular set of errors should not be incremental but rather complete (in a way that will become precise). The core argument here will be that – regardless of which kind of grammar is chosen – gradual learning should not be derived from gradual re-ranking.

Two tests of gradual learners are considered in this paper: the intermediate stages the learner passes through, and the end-state grammars that it eventually chooses. The first comes from a particular class of attested developmental stages, one in which children are preferentially faithful to material in the privileged positions of target words (which will be assumed here to be synonymous with inputs). In this paper I will use the term ‘Intermediate Faith’ (IF) as a cover term for this class of stages, and analyze them with crucial reference to positional faithfulness constraints. Under the standard assumption that initial-state phonological grammars allow little markedness in outputs and eventually come to allow all the marked structures on the target (Smolensky, 1996 *et seq*), the position of an IF stage along the learning trajectory is shown schematically below:

(1) Three stages of constraint re-ranking in development

- | | | |
|----|---------------|--|
| a. | initial state | $*X \gg \text{Faith-X(PrivilegedPosition)}, \text{Faith-X}$ |
| b. | intermediate | $\text{Faith-X(PrivilegedPosition)} \gg *X \gg \text{Faith-X}$ |
| c. | target | $\text{Faith-X(PrivilegedPosition)}, \text{Faith-X} \gg *X$ |

The literature contains many examples of children whose phonologies contain IF stages; section 2 will present two examples from this perspective.

The learnability question is what kind of constraint-based learner can be induced to pass through an IF stage like (1b) on its way from (1a) to (1c). The reason this particular stage is a challenge for a gradual learner that re-ranks gradually, using each error to affect a slight re-ranking, is that such stages cannot be reproduced by an OT learner directly from the frequency of constraint violations. This claim will be spelled out more in section 3, but the basic idea is as follows. Suppose the learner is being exposed to a language that tolerates complex onsets and has word-final stress. In the initial state, the high-ranking markedness constraint $*\text{COMPLEX}$ will ensure that complex onsets will be simplified at the expense of faithfulness. If we suppose that our learner’s constraint set includes two faithfulness constraints against deletion – general MAX and more specific $\text{MAX/STRESSED SYLLABLE}$ – then the learner will be making errors with violation profiles shown below:

(2) Two kinds of complex onset errors at the initial state:

	<i>Target output</i>	<i>Learner’s output</i>	<i>Constraints violated by the error</i>
a. deletion in stressed σ :	CV.'CCV	CV.'CV	MAX, MAX/STRESSED
b. deletion in unstressed σ :	CCV.'CV	CV.'CV	MAX

What the final column of this table shows is that a more general faithfulness constraint is violated *more frequently* than a specific one. This means that if frequency directly drives the rate at which rankings are revised by the gradual learner, the general MAX constraint’s importance in the ranking will be increased *faster* than that of the specific MAX/STRESSED . Thus, a purely frequency-driven gradual learner will not reach an IF stage where complex onsets are protected only in stressed syllables, under the influence of MAX/STRESSED . Nevertheless, section 2.1 points to a stage of French acquisition in which just this pattern is attested.

The direct correlation between frequency of violation and gradual re-ranking, shown to cause problems in the example above, is perhaps best known from the Gradual Learning Algorithm (GLA: Boersma, 1997; Boersma and Hayes, 2001¹)—though of course many authors have used more nuanced versions of a GLA learner.² Section 3 presents the core of the GLA, the stochastic OT grammars it learns and its problem with IF stages, setting the stage for the rest of the paper’s investigation.

In section 4, I provide my own proposal of an alternative method of gradual learning. The key idea is to begin with a very efficient, non-gradual OT learning algorithm – one that knows nothing of violation frequencies – and then to gradually feed it errors that will each push it to the next developmental stage. This approach is called Error-Selective Learning (ESL: Tessier, 2006, 2007), and its constraint re-ranking algorithm is a version of Biased Constraint Demotion (Prince and Tesar, 2004), which also incorporates aspects of the Low Faithfulness Constraint Demotion (Hayes, 2004). Section 4 demonstrates how ESL will produce IF stages if the right errors are available to the learner, illustrating with the French example already schematized above.

Section 5 presents a recent alternative, quite different approach to IF stages: one in which the a method very similar to the GLA is retained as the method of gradual learning, but where the method of evaluation is changed: replacing OT’s constraint ranking with constraint *weighting*, as in Harmonic Grammar (Legendre et al., 1990a,b, 2006; Pater, in press; Pater et al., 2007a,b; see also Prince, 2002a).³ Following Jesney and Tessier (2007, in press), section 5 shows how an incremental, GLA-like learner of a weighted constraint grammar⁴ will pass through the Intermediate Faith stage in (1b), purely as a result of the way its evaluation metric chooses optimal forms. The Harmonic Grammar approach is presented here because it demonstrates the need for further evidence to distinguish between the gradual re-ranking and gradual error-accumulation approaches.

With these two possible approaches to IF stages in mind, the paper then turns to the second testing ground: a gradual learner’s ability to converge on the correct end-state grammar. Here I argue that despite the success of the GLA-like learner at producing IF stages in a Harmonic Grammar, it still has drawbacks as a model of gradual human phonological development, and that these stem from its reliance on frequency of violation and incremental re-ranking. Section 6 presents a case in point, which I will refer to as the ‘winner misparse’ problem. The crucial initial observation is that before any learner can correctly use its errors and their violation frequencies to learn, it must have made the correct representational assumptions about the hidden structure of those errors—syllable boundaries, foot groupings, morphological structure, and so forth (e.g. Tesar, 1998; see also Dresner, 1999). Should the learner make the wrong assumptions temporarily, the learner may fall into a so-called ‘superset language’ (Berwick, 1985 *et seq*), from which the learner must recover in order to reach the correct end-state grammar.

As is spelled out in section 6, the relevance of this superset trap is that recovery to a more restrictive final grammar is unproblematic for the Error-Selective Learner, falling out rather directly from the mechanisms by which gradual development is achieved, but is much more problematic for a gradual re-ranker like the GLA. I present two hypothetical examples to make this argument, emphasizing that this issue arises through the interactions of both faithfulness and markedness constraints, and using either ranked or weighted constraints. The discussion of these ‘winner misparse’ traps also emphasizes the need to use *stored* errors in gradual learning, rather than a purely online learner that remembers only one error at a time.

Section 7 concludes the paper, including a summary of the various constraint-based learners presented, and also discusses how the type of grammar adopted (constraint ranks vs. weights) and the type of learning algorithm adopted each contribute to a theory’s success in learning restrictively and yet gradually.

¹ See also Apoussidou (2007), Boersma and Apoussidou (2004), Boersma and Levelt (2000), Boersma and Hayes (2001), Curtin and Zuraw (2001) and Levelt and van de Vijver (2004).

² Including but certainly not limited to Hayes and Londe (2006) and Zuraw (2000). See also section 6.

³ There are currently a number of independent approaches to learning phonologies with weighted constraints, such as Maximum Entropy (Goldwater and Johanson, 2003; Jaeger, in press), as well as OT grammars which themselves are weighted, using Maximum Likelihood estimation (Jarosz, 2006). None of these will be explored in any detail in this paper, though their ultimate connection to the issues raised here is an important topic for future research; however, see the arguments of section 7.2, which apply in principle to any grammar of weighted constraints. For some related but different comparative results, see Jesney (2007).

⁴ Note that this learner is very similar to the GLA, but not strictly speaking quite the same as the GLA: on the difference and the connection to the perceptron update rule of Rosenblatt (1956), see Jaeger (in press), Pater (2008a) and Jesney and Tessier (2007). In an attempt to minimize a potentially infuriating terminological confusion on this point, I will always use the term ‘GLA-like algorithm’ when discussing the learning of Harmonic Grammars in section 5, but readers not invested in the details may pretend this to be the familiar GLA algorithm without losing any of the plot.

The structural constraint responsible for the general reduction of clusters across stages can be defined simply as in (5):

(5) *COMPLEX: No tautosyllabic consonant clusters

Following a number of previous studies, I will assume that the constraint responsible for (4b)'s faithfulness in stressed syllables is a positional faithfulness constraints as in (6a) below. This constraint protects segments in the perceptually salient context of input stress, defined here over the syllable. (For discussion of such constraints in various analyses see, e.g. Beckman, 1998; Smith, 2001, as well as Curtin's, 2002 MAX-PITCHPROM; cf. Steriade, 1999). I also adopt the usual MAX constraint in (6b) relevant to all segments, regardless of stressed quality⁶:

(6) a) MAX-IO/STRESSED SYLLABLE

An Input segment in a stressed syllable must have an Output correspondent

b) MAX-IO:

An Input segment must have an Output correspondent

The ranking that accounts for the Intermediate Faith stage is one that sandwiches *COMPLEX between the two faithfulness constraints, as in (7). This grammar produces the right results, as illustrated in (8)⁷:

(7) The French IF stage: MAX/STRESSED >> *COMPLEX >> MAX

(8) Effects of the French IF stage

a. MAX/STRESSED >> *COMPLEX protects clusters in stressed syllables

/glɪs/	MAX/STRESSED	*COMPLEX
klɪs		*
kɪs	*!	

b. *COMPLEX >> MAX reduces clusters elsewhere

/gli.'sɑd/	MAX/ STRESSED	*COMPLEX	MAX
kla.'sæd		*!	
ka.'sæd			*

Some similar French data comes from Kehoe and Hilaire-Debove (2004), with respect to the acquisition of consonant-glide rather than stop-liquid sequences (though see that work for the author's interpretation of their results). The 14 children in their experiment (ages 1;10–2;9, mean age 2;4) preserved both members of two consonant-glide clusters (consonant-[w] and consonant-[ʃ]) more often in stressed than unstressed syllables ($p < 0.01$). For two children, the effect was fairly categorical, in that consonant-glide sequences were retained 100% of the time in stressed syllables, but less than 20% of the time in unstressed syllables.

⁶ This definition assumes that inputs are syllabified, at least for children/learners. For the purposes of this paper I will make this assumption without justification, although ultimately one must be found. For some summary of the known issues see Tessier (2007:102–104); I take these definitional concerns to be an important and unresolved issue. In this case, the constraint Max-PitchProm might be defined carefully so as to include all and only the onset segments adjacent to a stressed vowel, but the details of such an account will be left aside here. An alternative analysis is to assume two markedness constraints, *COMPLEX/UNSTRESSED SYLLABLE and *COMPLEX, and a single MAX constraint. As alluded to at the top of this section, I do not argue explicitly against such an alternative in this paper; I only refer to the body of work showing that *some* positional faithfulness constraints must exist. In that light, the learner will inevitably be faced with constraint pairs similar to those in (6a) and (6b), and thus will need to manage their consequences.

⁷ For evidence of this pattern in adult language, see Goad and Rose (2004) on Brazilian Portuguese.

2.2. Intermediate Faith to English stressed syllables

A different kind of IF stage is found in the extensive literature on syllable truncation, where it is often found that children resist the pressure to delete entire syllables from a privileged position (see, e.g. Echols and Newport, 1992; Fikkert, 1994a,b; Gerken, 1996; Pater, 1997). With respect to the stressed syllable position, Kehoe and Stoel-Gammon (1997) and Kehoe (2000) report on an elicitation study of English-speaking children at 2;4 and 2;10, designed in part to test for stress effects on syllable truncation. In their data, truncation patterns were almost exclusively restricted to unstressed syllables while stressed ones were retained.

A good example of stressed syllable preservation that suggests a role for MAX/STRESSED comes from one child in this study, 27m6. From the present perspective, this child's truncation patterns fall into two categories. In word with one input stress, outputs are truncated to a single trochaic foot (data all taken from Kehoe, 2000: Table 6).

9) 27m6's productions of target words with one stress

Schema:	Examples	
/wS/ → (S)	'giraffe'	[dwæf]
/wSw/ → (Sw)	'banana'	['bani]
	'tomato'	['medo]
/Sww/ → (Sw)	'elephant'	['ɛlɪf]

The examples in (9) show that at this stage, 27m6 always retains the target stressed *vowel*; the onset that surfaces in the output stressed syllable may also be the target stressed syllable vowel (as in 'tomato') or another input consonant (as in 'banana' and possibly 'giraffe'). The grammar's choice of unstressed target onsets in these latter cases can be attributed to markedness constraints that independently prefer low-sonority onsets (see more on this later in the section).

For words with two input stresses, both stresses must be retained in some form. This can be seen in 27m6's spontaneous productions of three and four syllable words below⁸:

10) 27m6's productions of target words with two stresses

Schema:	Examples	
/SS/ → (S)(S)	'raccoon'	[,ræ'kun]
/SwS/ → (Sw)(S)	'kangaroo'	[kæŋno'ja]
/SwSw/ → (Sw)(S)	'alligator'	[ˈæbrɪgɛɪ]
	'helicopter'	[ˈhɛkəpɔː]
(S)(Sw)	'avocado'	[a'kado]

To explain the one-foot pattern of truncation in (9), we can use markedness constraints that require a foot to be aligned with both the left and right edges of the word. This one-foot stage in phonological development has often been derived using gradient Align (Ft, PWd) constraints (e.g. Pater, 1997), but in light of McCarthy's (2003) influential arguments against such gradient constraints an alternative analysis is called for.⁹ In fact, McCarthy (2003) re-interprets single foot (i.e. non-iterative) stress systems in adult grammars as the result of END-RULE constraints (following Prince, 1983). These require the head foot in a Prosodic Word to be either the first or the last:

11) END-RULE-LEFT/RIGHT (Prince, 1983; McCarthy, 2003: p. 111)

The head foot is not preceded/followed by another foot within the Prosodic Word

⁸ One additional /SwS/ word, 'dinosaur', fluctuated in imitations between SS and SwS productions.

⁹ Thanks to an anonymous reviewer for pushing me to pursue different analyses of this stage.

If *both* of these constraints are ranked high, along with the constraint PARSE- σ which requires all syllables to be footed, all outputs will be truncated to a single foot. This ranking, given in (12), results in truncation of unstressed syllables that do not form part of the head foot, both at the beginning and ends of words, illustrated in (13):

12) Truncation to a single foot as in (9): PARSE- σ , END-RULE L/R >> MAX

13a) /wSw/ truncated to (Sw)

/ ^l ɛlɪfənt/	PARSE- σ	END-RULE-R	MAX
(^l ɛlɪ)fənt	*!		
(^l ɛlɪ)(fənt)		*!	
☞ (^l ɛlɪf)			***

13b) /Swɔw/ truncated to (Sw)

/bənæni/	PARSE- σ	END-RULE-L	MAX
bə(^l nani)	*!		
(bə)(^l nani)		*!	
☞ (^l bani)			***

Despite this general pattern, the longer words in (10) show that END-RULE L/R are in fact violable in this grammar—just in cases where truncating to a single foot would require deletion of a *stressed* syllable. This means that the same MAX/STRESSED constraint from the analysis of French is also crucial here: ranked above the END-RULE constraints, it derives the right results as in (14).¹⁰

14) Preservation of two feet in (15): MAX/STRESSED >> END-RULE L/R

15) a) /SwS/ preserved faithfully:

/kæŋ.gə. ^l ju/	MAX/STRESSED	END-RULE-L
(^l jɑ)	***!	
(^l kæŋno)	**!	
☞ (kæŋno)(^l jɑ)		*

15) b) /SwSw/ targets retain two feet:

i) primary stress on the initial foot: *helicopter*

/ ^l hɛlɪkaptər/	MAX/STRESSED	END-RULE-R
(^l kapə)	**!	
☞ (^l hɑ)(kapə)		*

ii) primary stress on the final foot: *avocado*

/ ^l avə ^l kado/	MAX/STRESSED	END-RULE-L
(^l kado)	*!	
☞ (ɑ)(^l kado)		*

¹⁰ Given the rejection of gradient alignment constraints above, it is interesting to note that four syllable words still lose one of their unstressed syllables. If both stressed syllables must be preserved, and there are no gradient alignment constraints to count the number of syllables beyond any foot edge: why should ‘alligator’ and ‘avocado’ suffer syllable truncation at all? In fact, this type of problem has already been addressed in the literature on adult stress systems: Gouskova (2003) uses structural constraints on foot well-formedness such as STRESS-TO-WEIGHT to analyze metrical syncope in languages such as Tonkawa and Southeastern Tepehuan, which look rather similar to this English child’s pattern. From this perspective, it may be that this latter stage of medial truncation reflects a desire for H rather than LL trochees, e.g. (hɑ_μ) rather than (hɛ_μ.li_μ) in Tableau (15b)i.

The grammar of 27m6 is thus another IF stage—one in which multiple feet are allowed, in violation of END-RULE, only when input stressed syllables are at stake.

16) Full ranking for English IF stage

PARSE- σ , MAX/STRESSED \gg END-RULE-L, R \gg MAX

A reviewer raises an alternative explanation for the truncation of unstressed syllables in developing grammars: that children have not encoded these unstressed syllables into their underlying representations, due in one way or another to their lesser salience. The evidence that input deficits cannot be the whole story comes from examples like *banana*, produced as [bani] in (13b). While 27m6 does not retain the entire unstressed syllable, his grammar nevertheless retains the unstressed syllable's onset segment, [b], reflecting the common tendency for developing grammars to select low-sonority onset segments. A more dramatic example of this comes from Gnanadesikan (1995/2004), who reports a child whose initial pretonic syllables were all over-written with a fixed segmental template [fi] (e.g. [fi.be.ya] for *umbrella*) but for whom the segmental content of the stressed syllable was similarly affected by the sonority of the unstressed syllable she had over-written. As in [bani], this child's grammar chose stop onsets over liquids, resulting in forms such as [fi.pis] for *police* and [fi.bet] for *barrette*. I take these patterns as evidence that unstressed syllable truncation must (at least sometimes) come from the input to output mapping, rather than defective input representations, and thus that the grammar adopted to explain truncation in (12) and (13) is a plausible one.

This section has presented two analyses of phonological development in natural language learning, each built around what I have called an Intermediate Faith ranking. In each case, learners present evidence of having acquired a marked structure – complex clusters in French, multiply-footed words in English – only in the privileged position of stressed syllables. The next three sections consider how a gradual constraint-based learner might derive these IF developmental stages.

3. Frequency of violation and the problem of IF stages

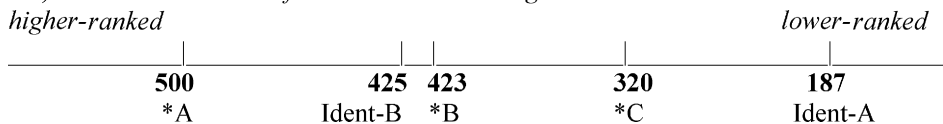
The first learner to be considered here is the basic GLA approach (Boersma, 1997; Boersma and Hayes, 2001). As already laid out in the introduction, this learner will not pass through an IF stage like the ones discussed in section 2, at least not unassisted. Yet the GLA is a powerful approach to learning with ambitious goals, and it is appealing in many respects—as such, the purpose of beginning with the GLA is not to prove it wrong, but to understand why and where it can go awry. I suggest that the problem is the particular way that the GLA relies on errors and violations in its gradual learning, and this diagnosis will lead to the alternative proposal in section 4.

3.1. Stochastic OT

The original view of OT as laid out by Prince and Smolensky (1993/2004) considers constraint rankings to be *ordinal*. Two constraints in such an OT grammar can only stand in one of two relations – $A \gg B$ or $B \gg A$ – and there is no sense in which A can be ranked *more or less* above B or vice versa.¹¹

In contrast, the constraint rankings that the GLA learns are what I will call *numerical* rather than ordinal. The GLA learns rankings where constraints have ranking *values* along a number line, so that every constraint is ranked not just above or below every other, but at a certain distance above or below every other. This is shown below, for some hypothetical constraints and ranking values:

17) *The numerical view of OT constraint ranking*



¹¹ In Tesar and Smolensky (1998, 2000) view of learning, this property is relaxed slightly to allow a third relation of *equal* ranking, at least during the course of learning. Thus, the initial state constraint ranking assumed in much recent work is of the form $\{M\} \gg \{F\}$, where all markedness constraints are ranked *above* all faith constraints, but ranked equally with respect to each other.

Furthermore, the grammar used by the classic GLA learner uses constraint rankings that are *stochastic*—they include some statistical noise. This noise is introduced by assuming that a constraint’s ranking value does not just represent its single point on the scale, as in (17), but rather the midpoint of a normal (i.e. Gaussian) distribution of values. In other words: constraint X’s ranking value is the place in the hierarchy that X is the most likely to sit, and the further away from X’s ranking value you get, the less likely X is to have that value. Each time a stochastic OT grammar is used, a *single* value is chosen for each constraint from its distribution of values—this choice creates a scale of single-point constraint values as in (18) below, which for practical purposes can be used by EVAL as a classic OT ranking:

18) *A one-time ranking*

502		424.7	422.95	320.078	184.342
*A		*B	Ident-B	*C	Ident-A

19) *The ordinal version of 18)*

*A >> *B >> Ident-B >> *C >> Ident-A

Despite the fact that each run of this grammar relies on a single ranking that can be equated with a classic OT hierarchy, there are crucial differences between ordinal and numerical OT. The ranking values in (17) show that IDENT-B is ranked above *B, but only slightly above; this means that their distribution of values overlap considerably. In the one-time ranking of (18), for example, the value chosen from *B’s distribution is in fact *higher* than the one chosen from IDENT-B’s, so that for this use of the grammar, their ranking has been reversed.¹² It is in this way that the relative distance between constraints makes numerical, stochastic OT different from the classic theory. It is also the conception of ranking values as numbers on a line that makes the Gradual part of the GLA possible, as we will see in the next section.

3.2. How the GLA learns its stochastic OT grammar

Like all other learning algorithms to be considered here, the GLA is error-driven. This means that it uses its current grammar to process language data and make errors; that it is the making of an error that triggers learning; and that the error guides it to reorganize its grammar in some way. To understand how the GLA learns, we must understand the format of these errors.

An error is an optimal candidate under the learner’s current grammar that is not identical to the observed (i.e. heard) winner. As an example, imagine that the learner provides the input /A/ to their current grammar, and EVAL returns the output [C]. The current grammar has thus made an error, illustrated in the tableau in (20):

20) *An error*

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

Our learner’s specific task to establish why it made an error – that is, why its current grammar mapped /A/ to [C], and not to [A] – so we can ignore the rest of the candidate set and just compare the two output candidates [A] and [C]. One way to make this comparison is the distilled form shown in (21):

21) *Boiling down the information in tableau 20)*

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

¹² The amount to which the curves of two constraints appear to overlap is a function not only of how similar their ranking values are but also how much random noise the system uses to choose one-time values: see Boersma (1997).

In Prince (2002a), this distillation of candidate comparisons is called an Elementary Ranking Condition vector; here I will refer to them as ERC rows. What each cell in an ERC row reports is the preference of each constraint with respect to the winner and its rival loser candidate. The tableau in (21) shows that *A assigns a violation mark to the winner [A], and no mark to the loser [C], so we can say that *A *prefers the loser*: thus the ERC row for the A ~ C comparison contains an L in the *A column. Similarly, the third markedness constraint *B assigns equal violation marks (in this case, none) to both the winner and loser candidates: thus, *it prefers both winner and loser equally*, and this equality puts an e in the *B column.

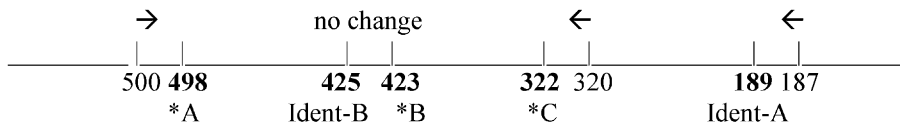
The Ls, Ws and es of an ERC row indicate the relevant discrepancies between the current and target grammars. The GLA's procedure of learning from these discrepancies is easy to describe: it promotes all constraints that prefer the winner (i.e. that assign a W in that error's ERC row) and demotes all constraints that prefer the loser (i.e. assign an L).¹³ So in response to the error in (21), the GLA will now adjust ranking values as follows:

22) *The GLA's response to the errors*

winner ~ loser	*A	*B	*C	IDENT-A	IDENT-B
A ~ C	L→	(no change)	←W	←W	(no change)

The amount by which each constraint is moved in response to an error is referred to as the learner's *plasticity*. If for example the learner's plasticity is currently 2, the actual re-ranking effect of (22) applied to the old grammar from (17) will be as in (23) below (previous ranking values are in regular font, new ones are in bold):

23) *The new GLA grammar*



Note that the GLA learner does not attempt to *resolve* errors in any immediate way: the grammar in (23) is only very slightly less likely to map /A/ unfaithfully onto [C] as the previous grammar was. As Boersma and Hayes (2001) put it:

“The hypothesis behind the Gradual Learning Algorithm is that moderate adjustments of ranking values will ultimately achieve the right grammar.” (p. 52)

... and that the extent of these gradual adjustments, over time, should mirror the frequency with which they have assigned Ws and Ls.

3.3. *The problem with IF stages*

This section walks through the way that frequency of violation prevents the GLA from reaching the IF stage. The example concerns the French complex onsets from section 2.1, as they most easily illustrate the problem.

We start with the initial state—the ranking that characterizes the state of the grammar before any learning has taken place. The basic assumption of nearly all the relevant learning literature is that the initial OT state is Markedness ≫ Faithfulness (see, among many others, Smolensky, 1996; Demuth, 1996; Gnanadesikan, 1995/2004).

¹³ This particular method of choosing constraints to promote and demote is really only one of many GLAs considered in, e.g. Boersma (1997) and Boersma and Hayes (2001). However, this is the one that these authors find works best—in particular, Boersma and Hayes (2001) diagnose this brand of GLA as the only one that produces the variation patterns they attempt to model. Therefore, I will refer to this re-ranking algorithm as “the GLA”.

The French IF stage is one seen halfway through development; in the target grammar the ranking of the relevant constraints is the reverse of the initial state: all $F \gg M$. Thus, we have the three rankings below¹⁴:

24) Three rankings

- a. *Initial state*: *COMPLEX \gg MAX/STRESSED, MAX
 b. *IF state*: MAX/STRESSED \gg *COMPLEX \gg MAX
 c. *Target state*: MAX \gg *COMPLEX

As indicated by the ranking in (24c), the target French stage is one that tolerates complex onsets in both stressed and unstressed syllables (recall Table 4). The following is a schematic illustration of how this kind of lexicon will prevent the GLA, as seen so far, from reaching the IF stage:

25) GLA learning, with schematic French

a. Hypothetical observed winners:

bá blá biblá blibá

b. GLA Constraint movement, created by errors at initial state:

*COMPLEX: demoted by every word with *any* complex onset, i.e:
 blá biblá blibá

MAX: promoted by every word with *any* complex onset, i.e.:
 blá biblá blibá

MAX/STRESSED: promoted by every word with *a stressed* complex onset, i.e:
 blá biblá

The upshot of (25b) is that no matter what the frequencies of word types, *COMPLEX and general MAX will fall and rise at the same rate during learning while specific MAX/STRESSED will rise *slower*: general MAX has the additional pressure of complex onsets deleted from unstressed syllables (*blabá*). Thus the second stage that this learner will reach is the one in (26):

26) The GLA's first new stage: MAX \gg *COMPLEX \gg MAX/STRESSED

Unfortunately, this is already the target stage: one that saves *all* complex onsets.

27) The mismatch between 43) and the observed intermediate stage:

	/blablá/	MAX/STRESSED	*COMPLEX	MAX
	(i) babá	**!		**
<i>IF stage winner</i>	(ii) bablá	*!	*	*
<i>GLA-learned winner</i>	(ii) [☞] blablá		**	

The core of the problem is this. When two faithfulness constraints are in a specific-to-general relationship like MAX/STRESSED- σ and MAX, the *frequency* with which they demonstrate their need to be promoted will never push the more specific one up *faster* than the more general one. Yet to reach the IF stage by gradually changing ranking values, this is precisely what must happen.

3.3.1. An alternative: no context-free faithfulness constraints?

A reviewer points out that a GLA learner would receive a different set of violation profiles if its constraint set included two complementary faithfulness constraints: one targeting stressed segments and the other unstressed segments. In this alternative, the learner will be faced with two sets of errors, each supporting the promotion of the one of the faithfulness constraints:

¹⁴ I assume that the ranking of the more specific faith constraint MAX/STRESSED is irrelevant here, although this is not always true: see Prince (1997) and Lombardi (1999).

28) GLA learning of schematic French, with a revised constraint set

a. Initial state: *COMPLEX >> MAX-IO/STRESSED, MAX-IO/UNSTRESSED

b. Hypothetical observed winners:
 bá blá biblá blibá

c. GLA constraint movement, created by errors at initial state:

(i) *COMPLEX: demoted by every word with *any* complex onset:
 blá biblá blibá

(ii) MAX/STRESSED: promoted by every word with *stressed* complex onset:
 blá biblá

(iii) MAX/UNSTRESSED: promoted by every word with *an unstressed* complex onset:
 blibá

As (28) shows, the errors that promote these two positional MAX constraints are entirely independent. What will determine which constraint the GLA learner promotes above *COMPLEX first, then, is purely a function of error frequency: that is, whether more target words are like (28bii) or like (28biii). Under the fairly reasonable assumption that French child-directed speech provides more stressed complex onsets than unstressed ones, our GLA learner could indeed reach the intermediate stage from section 2.1, as shown in (29):

29) The intermediate stage using revised constraint set:

MAX/STRESSED >> *COMPLEX >> MAX/UNSTRESSED

a)

/biblá/	MAX/STRESSED	*COMPLEX	MAX/UNSTRESSED
(i) ☞ biblá		*	
(ii) bibá	*!		

b)

/blibá/	MAX/STRESSED	*COMPLEX	MAX/UNSTRESSED
(i) blibá		*!	
(ii) ☞ bibá			*

While this constraint can allow the learner to reach the desired intermediate stage, it also allows the possibility of the reverse pattern, in which MAX-UNSTRESSED is highest ranked and complex onsets are allowed only in *unstressed* syllables. Since such a pattern is attested neither in the developing grammars of children, nor in the typology of cross-linguistic adult grammars, it would seem that revising the constraint set into mutually exclusive specific faith alternatives is unsatisfactory as a general solution to the learning of intermediate stages.

To re-emphasize the main point, the question raised here is how closely tied the gradual learner should be to the frequency of the errors and violations it encounters. It is both well-documented and unsurprising that the ambient frequency of some phonological structures affects the course and rate of their acquisition: for some pertinent examples, see Curtin and Zuraw (2001), Levelt and van de Vijver (2004), Roark and Demuth (2000) and Stites et al. (2004). So the rest of this paper asks: how might a learner be gradual and influenced by violation frequencies, but still pass naturally through IF stages?

Before continuing, a sidenote on alternatives. A more elaborated version of the GLA could well be used to avoid problems caused by subset-superset relationships between faithfulness constraints such as this. The concrete proposal would be to equip the learner with *three* initial constraint rankings, with markedness very high, specific faithfulness in the middle, and general faithfulness very low. However, Hayes and Londe (2006: section 6.5) report that using a GLA learner alone is insufficient to restrictively capture the grammar of Hungarian vowel harmony, and they cite in particular a general faithfulness constraint (IDENT-IO/BACK) that ‘inevitably rises too high in the grammar.’ (See also Tessier, 2007 for further discussion of difficulties for this option). In section 6, I will return to the Gradual Learning Algorithm and its variants (albeit with a different grammatical framework) and argue that they are still insufficient to avoid superset language traps. But since the larger argument here is that our gradual learner should *not* proceed via gradual-reranking, I will first present my alternative proposal below.

4. The Error-Selective Learning approach to IF stages

This section presents the Error-Selective Learner.¹⁵ In contrast to the previous approach, the ESL idea is to use a learning algorithm that knows nothing about frequency or gradualness, and to instead build these sensitivities into the way it chooses errors to learn from.

4.1. Background on Error-Selective Learning

The Error-Selective Learner is built around an algorithm that, unlike the GLA, uses ordinal OT rankings. This algorithm is a blend of two recent and influential proposals: Biased Constraint Demotion (BCD: Prince and Tesar, 2004) and Low-Faithfulness Constraint Demotion (LFCD: Hayes, 2004). In what follows, I adopt much of the terminology of BCD, but I also rely crucially on a Faithfulness bias proposed only in LFCD. I will refer to this particular amalgamation as Multiply-Biased Constraint Demotion (MBCD).¹⁶

To repeat the starting point: the learner with a grammar like the one in (30a) will make errors like the one represented in (30b):

30) Creating errors

a. An ordinal constraint ranking: *A >> *B >> IDENT-B >> *C >> IDENT-A

b. An ERC row created by using a) to map /A/ --> [C]:

/A/	*A	*C	*B	IDENT-A	IDENT-B
A ~ C	L	W	e	L	e

Given an ordinal view of OT grammar as in (30a), learning cannot be *gradual* re-ranking of constraints, up and down a GLA-style number line: instead, learning must mean changing the *order* of constraints. The goal of re-ranking for this learner is to ‘resolve’ the error—that is, to build a new ranking in which the produced loser [C] is less optimal than the target form [A]. Each cycle of learning creates a new grammar hypothesis, and this new grammar will cause a new set of errors and consequent ERC rows. While previous grammars are forgotten as soon as a new one is built, this learner *retains* its ERC rows in a table called the Support—thus, (30b) could be one of the many ERC rows in the learner’s Support at one point in learning. An important aspect of the Error-Selective Learner is that its MBCD algorithm always works with reference to the Support.

The logic used to resolve errors here comes from the Cancellation/Domination Lemma of Prince and Smolensky (1993:148); described in Prince and Tesar (2004:255) like this:

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

This lemma is the crux of the recursive Constraint Demotion Algorithm (CDA: Tesar and Smolensky, 1996, 1998, 2000; see also Prince, 2002a,b) and it also drives the core of MBCD and the algorithms it is based on. To rephrase the lemma a little: this algorithm will ensure that [A] is a more optimal output than [C] by installing constraints that are better satisfied by /A/ → [A] above constraints that are better satisfied by /A/ → [C].

Building from the CDA, the re-ranking algorithms in Prince and Tesar (2004) and Hayes (2004) were designed with a particular goal: to be *restrictive*. To be maximally restrictive, a learner must choose a grammar that faithfully reproduces all the attested forms, and allows as few *other* forms as the constraint set allows.¹⁷ This is by no means an easy task—at the very least because many different constraint rankings will choose the same optimal input for a given output, and that each ERC row will only partially determine the nature of the new grammar to be learned.¹⁸ Space constraints do not permit a

¹⁵ Error-Selective Learning is proposed in a somewhat different format in Tessier (2006, 2007).

¹⁶ I will leave aside here the issues of how other biases from LFCD, such as ‘Prefer Active’, must also be used by this type of learner.

¹⁷ Put somewhat differently: a properly restrictive grammar maps all of the Rich Base of potential inputs onto attested output forms.

¹⁸ An incomplete history of restrictive linguistic grammar-building, in other frameworks as well as OT, includes but is not limited to Angluin (1980), Berwick (1985), Dresher (1999), Dresher and Kaye (1990), Gibson and Wexler (1994), Hayes (2004), Ito and Mester (1999), Jarosz (2006), McCarthy (1998), Prince and Tesar (2004), Pulleyblank and Turkel (1998), Smith (2000), Smolensky (1996), Tesar and Smolensky (1996, 1998, 2000) and Tessier (2006, 2007).

full explanation of how BCD or LFCD ensures restrictiveness (see the original references for much more thorough discussion); the important ways that MBCD strives for restrictiveness will be introduced in section 4.2 as needed.

Since such algorithms resolve errors completely, they are not learners that go through any intermediate stages, such as the IF stages of section 2. This is not surprising: Prince and Tesar (2004) in particular are explicit in stating that their goal is not to model childhood acquisition, but rather to solve a formal learnability problem (namely how best to impose maximal restrictiveness on OT learning). Any time MBCD is applied to the Support, it will learn everything there is to learn from each error, and ensure that error is never made again. Therefore, a pure MBCD constraint-demotion learner cannot learn gradually.

ESL is a way of using such an efficient re-ranking algorithm to learn gradually. The basic idea is to use Multiply-Biased Constraint Demotion to do what it does well (i.e. choosing the correct constraint ranking given a Support) but to slowly feed the right errors to the Support. Given that MBCD ensures restrictiveness by imposing ranking biases, ESL uses these same biases to select the best errors to learn from. Thus, each ESL learning cycle proceeds in two core steps: first, choosing a set of potential errors to learn from, and then applying MBCD and its biases to this set of errors until some learning has taken place. The result is a slightly modified ranking, and one new error added to the Support¹⁹—and then the process begins again.

4.2. Deriving the IF stage of French cluster acquisition

In the ordinal OT view, the initial state of the French learning scenario has *COMPLEX ranked above both faithfulness constraints:

32) *The French initial state (fragment):* *COMPLEX >> MAX, MAX/STRESSED

The goal for our learner is to get from this initial state constraint ranking in (32) to the IF grammar from section 2.1.

4.2.1. Making and storing errors

As assumed throughout, the Error-Selective Learner begins by taking target forms as inputs and using its current grammar to produce outputs. However when this learner makes an error, it does not immediately add the error to the Support; neither does this error immediately trigger re-ranking. Instead, as errors are made, their resulting ERC rows are put into a temporary storage area called the ‘Error Cache’.

This is illustrated below in (33): one early French ranking, given in (a), produces errors like the ones in (b). In this example, I have added some additional plausible markedness constraints relevant to the particular target segments being produced unfaithfully in the outputs.²⁰ For the sake of space, these additional constraints have been abbreviated in the tables below, labeled with the particular segment that they penalize. Note that the dotted lines on the Error Cache table are meant to indicate the impermanence of this object.

33) a. *An early grammar fragment of French:*

{*COMPLEX, *VCDALVEOLARFRIC, *UVULAR, *VOICEDVELARSTOP, *FRONTROUNDV}
>>
{MAX-IO, MAX-IO/STRESSED}

33) b. *An Error Cache for the grammar in a)*

input	winner ~ loser	*COMPLEX	*z	*ʁ	*g	*y	MAX/ STRESSED	MAX
/gʁo/	gʁo ~ ko	L	e	L	L	e	W	W
/bʁi'ze/	bʁi'ze ~ pi'ze:	L	L	L	e	e	e	W
/gʁy.'jo/	gʁy.'jo ~ ku.'jo	L	e	L	L	L	e	W

While this grammar fragment is at the initial state with respect to these constraints, it is clearly the case that initial states of production do not necessarily correlate with a truly initial grammar—that is, some preliminary demotion of

¹⁹ Or at least a small set of errors—see section 4.2.

²⁰ In other words: each describes a class of segments missing from various languages: voiced velar stops are lacking in, e.g. Dutch (Booij, 1995), uvulars and front rounded vowels in, e.g. English, voiced fricatives in, e.g. Tagalog (Zuraw, 2000).

markedness constraints has already taken place (see esp. Velleman and Vihman, 2003). In the case of Théo—it seems reasonable to assume from the losers in (33b) that his grammar has demoted IAMB below TROCHEE to allow final stress, in addition to the rankings given in (33a).²¹ This means that his Support must already contain an error which brought about this demotion. To add this element of realism to this scenario, which will prove important to the workings of the Error-Selective Learner, we will assume that the Support currently contains some error indicating this fact—for example an error made on the word for ‘baby’, *bébé*:

34) The current Support, with one error from /be'be/ ‘baby’

<i>target</i>	<i>winner ~ loser</i>	TROCHEE	IAMB	*COMP	*z	*ɿ	*g	*y	MAX STRESSED	MAX
/be'be/	be'be ~ 'bebe	L	W	e	e	e	e	e	e	e

Due to the confines of the printed page, this Support row in (34) should be interpreted as a fragment of the full ERC row. While this word is assigned *es* by all the featural markedness constraints used in (33b) above, it will have triggered learning on other markedness constraints such as NONFINALITY, and so the current grammar will have demoted them along with TROCHEE accordingly.

4.2.2. Choosing an error to learn from²²

As the learner continues to use its current grammar, the errors pile up and the Cache grows, while the Support is not updated and retains only its single error as in (34). To get anywhere more learning must eventually occur, which means that re-ranking must be triggered. In the Error-Selective model, learning is triggered when a markedness constraint overcomes the ‘Violation Threshold’—that is: when some constraint has assigned an L to more than *x* number of words in the Error Cache.²³ This offending constraint is called the Trigger Constraint, because it has triggered learning. To take a very unrealistic but illustrative example: setting the violation threshold to 3 means that as soon as some markedness constraint assigns an L to three different winner ~ loser pairs in the Error Cache, learning occurs.

In the Cache of (33b) above, adding that last error on ‘oatmeal’ means that *COMPLEX has now become a Trigger Constraint, and so a learning cycle begins. Step One of ESL is for the learner to choose a set of what are called Potential Best ERCs, being those ERC rows that might be learned from. This work is done by a sub-routine called the Error-Selection Algorithm (ESA); since our goal here is gradual learning, the ESA is designed to choose errors that will require as *little* change to the current ranking as possible. The ESA is defined in (35):

35) ESL Step One: The Error Selection Algorithm

Choose as the Potential Best ERCs those rows in the Cache which:

- a) have an L assigned by the Trigger Constraint, and of those, the ones that**
- b) have the fewest Ls assigned by other Markedness constraints**

The two ESA criteria narrow down our set of three errors in the Cache of (33b) to two. All three errors have an L assigned by the Trigger Constraint *COMPLEX, so they all meet criterion (a). Criterion (b) rules out the last error ‘oatmeal’, since it has Ls assigned by *three* other markedness constraints; the other two errors remain as they each have Ls assigned by *two* other markedness constraints. Thus the ESA algorithm chooses our first two errors as the Potential

²¹ It seems to be anecdotally acknowledged that the position of word stress is one of the first thing that children acquire in their phonology: that children make few errors in phonological stress, and overcome them much sooner than, e.g. segmental or featural errors. This tendency will be derived if errors that demonstrate stress patterns like this one in (34) are added to the Support early on; why the ESL mechanisms are likely to choose stress errors earlier on is returned to in section 4.3.

²² I am grateful to an anonymous reviewer for critiques of this aspect of ESL, which I believe have improved the current version.

²³ This number is an independent parameter, whose setting is one way that this formal learning model might encode something like the amount of cognitive attention being paid to the learning problem. In other words: the lower the VT, the fewer errors the learner needs to be persuaded to learn a new grammar. See Tessier (2007) for some discussion.

Best ERCs set. (This is of course not an accident—the Cache in (34b) is designed to restrict this example’s focus just to errors with complex onsets in the two relevant positions):

36) *The Potential Best Errors, chosen by the ESA in Step One*

input	winner ~ loser	*COMPLEX Trigger	*z	*ʁ	*g	*y	MAX STRESS	MAX
/gʁo/	gʁo ~ ko	L	e	L	L	e	W	W
/bʁi'ze/	bʁi'ze ~ pi'ze:	L	L	L	e	e	e	W

The crucial difference between the two rows in (36) is in their faithfulness violations—in particular, whether the specific faithfulness constraint MAX/STRESSED assigns a W or not. In the first error ‘gros’, markedness has caused deletion of a *stressed* onset segment so that both MAX constraints assign a W; for the second error ‘brisé’, markedness has caused the deletion of an *unstressed* segment so the specific MAX constraint is assigned an e.

4.2.3. *Learning from errors using MBCD*

Step Two of the ESL cycle is to apply the MBCD algorithm to build a new grammar—including both the old Support AND the chosen set of Potential Best ERCs. The central idea of Multiply-Biased Constraint Demotion, like the BCD and LFCD algorithms it is built from, is to give the learner a set of constraint ranking biases, which the learner assumes up until the ERC rows provide evidence to the contrary. Building a constraint ranking is a series of cycles of adding constraints to strata—starting at the top and continuing until there are no more constraints to be ranked. In building each stratum, the learner aims to install all constraints that its biases want highest-ranked, and put off the installation of all other constraints until it has to.

The MBCD algorithm that I adopt here has two important ranking biases. First is the by-now-familiar markedness \gg faithfulness bias which makes the first pass of ranking decisions; second is a bias for ranking more specific IO-Faith constraints above more general ones, proposed by Smith (2000), and implemented in LFCD by Hayes (2004). Both of these are crucial to seeing how our learner with the Support in (51) will reach the IF stage grammar.²⁴

With these biases, the MBCD algorithm builds a grammar by performing a total ordering of the constraints from scratch, in a way that will resolve all the ERC rows it is given. In Step Two, the Error-Selective Learner’s goal is to use the MBCD algorithm’s ranking biases to install constraints, and thereby resolve errors in the Support, *up until the point that one Potential Best ERC row has been resolved*. We will see immediately how this short-term goal will ensure gradual learning of the sort that derives IF stages.

37) *ESL Step Two: Applying MBCD (first pass)*

- a) **Begin with the set of ERC rows that includes both (i) all ERC rows in the Support and (ii) all the Potential Best ERC rows.**
- b) **Apply MBCD to this set, installing constraints in strata until one of the ERC rows in (a) has been resolved**

In the present example, MBCD begins with the set of constraints from (33) and (34) above, all as yet unranked, and the errors in (38) below to resolve:

38) *Errors for the MBCD at the beginning of Step Two*

input	winner ~ loser	TROCHEE	IAMB	*COMP	*z	*ʁ	*g	*y	MAX STRESS	MAX
/be'be/	be'be ~ 'bebe	L	W	e	e	e	e	e	e	e
/gʁo/	gʁo ~ ko	e	e	L	e	L	L	e	W	W
/bʁi'ze/	bʁi'ze ~ pi'ze:	e	e	L	L	L	e	e	e	W

²⁴ Given the importance of the specific \gg general IO-faith bias here, it must be noted that Prince and Tesar (2004) present good reasons to explicitly reject such a bias, using different principles to determine which IO-faith constraints to install in any stratum. For reasons of space, I simply assume here that this bias can indeed be implemented; for the details see Hayes (2004) and Tessier (2007).

MBCD then proceeds to build each stratum in the constraint ranking as follows.²⁵

- Stratum 1: (i) install all the Markedness constraints *that prefer no losers*:
resulting stratum 1: *FRONTROUNDV, IAMB

These are the only constraints we can install under (i): the other markedness constraints each prefer at least one loser, and the last two are not markedness constraints. This first principle is the main way in which the $M \gg F$ bias is instantiated in MBCD.

Having installed *y has not yet resolved any errors (since it did not assign any Ws), but installing IAMB has. With this W-preferring constraint at the top of the hierarchy, it necessarily outranks all the L-preferring constraints in the ERC row for 'bébé', and so by (31) we now that this error has been resolved. Once an error has been resolved, the MBCD algorithm stops looking at that error's Ws and Ls:

39) Errors for the MBCD after Stratum 1 built:

input	winner ~ loser	Trochee	Iamb	*Comp	*z	*ɹ	*g	*y	Max Stressed	Max
/be ^h be/	be ^h be ~ be ^h be	⊥	W	e	e	e	e	e	e	e
/gɹo/	gɹo ~ ko	e	e	L	e	L	L	e	W	W
/bɹi ^h ze/	bɹi ^h ze ~ pi ^h ze	e	e	L	L	L	e	e	e	W

Now the algorithm moves onto the next stratum, beginning again with the first principle, and since the first error is now ignored we can install another markedness constraint:

- Stratum 2: (i) install all the Markedness constraints *that prefer no losers*:
resulting stratum 2: TROCHEE

And then we move on:

- Stratum 3: (i) install all the M constraints *that prefer no losers*
(but each prefers a loser)

Since (i) can install no constraints, MBCD moves onto the second decision principle, which has two parts:

- Stratum 3: (ii) find all the Faithfulness constraints *that prefer a winner*:
con'd (and there are two: MAX/STRESSED and MAX)
- (iii) install (one of) the W-preferring F-constraints *that is the most specific*
resulting stratum 3: MAX/STRESSED

The (ii) part of this principle is an extension of the $M \gg F$ bias—since we are forced to include a faithfulness constraint in this stratum, we should at least include one that will resolve some error so that the *next* stratum has a better chance of containing M constraints. The (iii) part imposes the second bias, Specific-Faith \gg General-Faith, and in this case it chooses MAX/STRESSED.²⁶

With just these two strata of constraints installed, we can now see that this new grammar will no longer make the error on 'gros' given in (38): MAX/STRESSED will rule out the loser for deleting a stressed segment in *['ko], so the

²⁵ Note that this walk-through is merely a synthesis of the core of BCD, augmented with a bias from LFCB, and certainly does not do justice to either.

²⁶ Looking at just these two constraints, it is obvious which one is more specific from their names and definitions, but this is not always the case: see Prince and Tesar (2004) and Tessier (2007).

faithful [ˈgro] will win. In the present terms, this ERC row has been resolved. And now our short-term Step Two goal has been met. We have built a ranking that resolved one of the Potential Best ERC rows; we have learned something new. With this learning accomplished, the Error-Selective Learner finishes Step Two by *removing* all the remaining Potential Best ERC rows from its learning datum except the one it has resolved.

40) *ESL Step Two: Applying MBCD (final)*

- a) Begin with the set of ERC rows that includes both (i) all ERC rows in the Support and (ii) all the Potential Best ERC rows.
- b) Apply MBCD to the set in (a), installing constraints in strata until one of the ERC rows in (ii) has been resolved
- c) Remove all of (ii) except the error resolved in (b)**
- d) Continue applying MBCD to the remaining Support ERC rows until termination**

41) *The new Support, after one Potential Best ERC row has been resolved*

input	winner ~ loser	Trochee	Iamb	*Comp	*z	*ɥ	*g	*y	Max Stressed	Max
/beˈbe/	beˈbe — bebe	⌊	⊘	e	e	e	e	e	e	e
/gʁo/	gʁo — ko	e	e	⌊	e	⌊	⌊	e	⊘	⊘

Since at this point all errors have been resolved, the MBCD algorithm is free to rank the remaining constraints as its biases please—i.e., all remaining $M \gg F$.

- Stratum 4 (i) install the Markedness constraints that *prefer no losers*:
resulting stratum 4: *COMPLEX, *UVULAR, *VCDFRIC, *VCDVELARSTOP
- Stratum 5: (i) install the Markedness constraints that *prefer no losers*:
(but each prefers a loser, so:)
- (ii) find all the Faithfulness constraints that *prefer a winner*:
(and there is only one remaining, MAX, so:)

resulting stratum 5: **MAX**

And since there are no more constraints left to be ranked, the MBCD algorithm has succeeded, and it terminates.

42) *The final result of applying MBCD to the error set in 38)*

```
{*FRONTROUNDV, IAMB}
  >>
{TROCHEE}
  >>
{MAX-STRESSED}
  >>
{*COMPLEX, *VCDFRIC, *UVULAR, *VCDVELARSTOP}
  >>
{MAX}
```

Having arrived at the grammar in (42), the Error-Selective Learner has reached the goal of this section. This ranking that MBCD has built includes the necessary rankings to produce the attested French IF stage with respect to complex onset simplification (see the bolded constraints). Compared to the previous grammar in (32), this ranking now protects complex clusters in stressed syllables, but still rules them out in unstressed ones.

At the end of Step Two, the learner now has a new ranking and a new error learned from. To keep track of the knowledge that moved it from the rankings (32) to (41), it must now update its Support, to include that ERC row that was resolved. In the present case, that was the ERC row for ‘*gros*’, resolved after Stratum 3 was built. Conversely, the learner must also clear their Cache of all the errors made by the previous grammar.²⁷ These two updates are Step Three of the Error-Selective Learning process; after this, the learner begins using their new grammar, making new errors that are accumulated in the Cache, and waiting until the Violation Threshold is overcome by a new constraint and the cycle begins again.

43) ESL Step Three: Updating Error Memory

- a) Add to the previous Support that ERC row which was resolved in Step Two (b)
- b) Empty the Cache

4.3. *Frequency of violation and Error-Selective Learning*

To take a step back: what we have just seen is that the Error-Selective Learner, using a non-gradual re-ranking algorithm, can nevertheless progress from the initial state to an IF stage. It does so by choosing to learn only from a small subset of errors that it has already made, and it uses its ranking biases in part to make those choices.

While the Error-Selective Learner is guided in all its learning decisions by violation profiles, its sensitivity to violation frequency is rather more complicated than that of the GLA—and its sensitivity to markedness vs. faithfulness violation frequencies is crucially different. Violation Thresholds, and the ESA criterion (b) that favours errors with the *fewest* other Ls, conspire to predict that order of acquisition should broadly mirror markedness violation frequency. The more errors that a markedness constraint assigns Ls to, the more of those L-assigned errors will pile up in the Cache, so the more likely it is to either be a Trigger Constraint, or to be one of the few Ls assigned to a members of the Potential Best Error set. Note that this sensitivity accords with the tendency discussed at the beginning of this section (see footnote 21): that the basic facts of phonological stress, such as the relative ranking of TROCHEE vs. IAMB, appear to be learned earlier across languages than featural markedness. Since every word has stress, but not every word necessarily contains a particular marked segment or featural combination, more Ls assigned by these basic metrical constraints will pile up in the Cache faster than for constraints like *FRONTROUNDVOWEL or *UVULAR, and so are more likely to overcome the Violation Threshold and/or be added to the Support.

On the faithfulness side, however, the MBCD ranking bias makes the choices. In fact, this biases the learner towards choosing errors with the *most* Ws assigned by faith constraints (compare the number of faith Ws assigned to the Potential Best ERCs in (36). The MBCD biases themselves choose which of the Potential Best Errors to resolve first, and they will choose to demote markedness below *specific* faith constraints first—even though the general faith constraints assign Ws more often.

It must be emphasized that while the Error-Selective Learner *can* go through an IF stage, it is by no means guaranteed to do so. This indeterminacy comes first from the Error-Selection Algorithm, which chooses the Potential Best ERC rows. In this example, the ESA chose errors with clusters in both stressed and unstressed syllables, precisely because each of these errors had the fewest *other* markedness problems given the then-current grammar. If a different error had been included in the Cache – one with an unstressed onset cluster and *fewer* assigned Ls otherwise – it would have been chosen as the lone Best ERC row, and its inclusion in the Support might well have derived a new grammar with the target ranking, MAX ≫ *COMPLEX. No IF stage would have been created in between.

Given this element of randomness, it should be noted as a positive sign that some IF stages are probably not shared by all learners of a language. This French example is a case in point, for while the two children in Rose’s (2000) study went through this IF stage, the Kehoe and Hilaire-Debove (2004) results provide evidence for a few children, going through the same stage with respect to two particular obstruent-glide clusters only, and no evidence of a positional asymmetry at all for the majority others. Determining whether some such stages are nearly universal within a particular language, and whether the ESL approach can predict that universality (from the constraint set and the input frequencies facts of that language) is an empirical and open question.

²⁷ For why the Cache must be cleared, see Tessier (2007).

4.4. Variation between rankings and Error-Selective Learning

An attentive reader, particularly one with restrictive grammar-building in mind, may have noticed something suspicious about the ranking in (42). What are the consequences of all the markedness constraints that were demoted down to Stratum 4, along with *COMPLEX? Some of these demotions are inescapable: once ‘gros’ has been added to the Support, all of the markedness that its loser eliminated must be allowed into the grammar. But why has the learner demoted the constraint against [z]? When learning began, all Potential Best Errors were at stake, so MBCD was prevented from installing *VCDALVFRIC in the top stratum by the error on ‘brisé’. But once ‘gros’ was resolved, ‘brisé’ and all the other PBEs were ignored, and subsequently cleared from the Cache. Thus is the final Support, there is no evidence for this constraint’s demotion! Does this mean that the ESL is doomed to be unrestrictive, despite the MBCD’s explicit goals?

Thankfully no—because ESL is crucially about storing errors, not grammars. In the ranking of (42), the learner has indeed built a grammar that allows more marked structures than it has stored evidence for. However, a central idea of Error-Selective Learning is that building rankings is always an easy task (getting from ERC rows to grammars via MBCD)—what the learner hesitates to do is build permanent representations of learning data (the ERCs themselves). If the MBCD algorithm were run *again* on the current Support, the learner will have no recourse to the previous Cache and its old errors, and so indeed will be able to install *VOICEDALVOLARFRICATIVE in the top stratum. Under these assumptions, then, the complete Error-Selective Learner will be required to run their current Support through MBCD—merely to clear out any vestigial rankings for which there is no stored evidence in its ERC rows. There are multiple reasons this might be a good idea beyond the current issue; I will return to another one in section 5.²⁸

A final remark about the nature of Error-Selective Learning: the goal of this approach is not to find maximally informative data for the learner to attend to—in comparison to, e.g. recent proposals by Pearl (2007, 2008). In fact, as a reviewer points out, the aim of ESL is in fact to select *uninformative* data, of a certain sort, so as to delay fully accurate learning of the target phonology.

5. The Constraint Weighting approach to IF stages²⁹

This section introduces a rather different answer to the question of how to use violation frequencies to reach an Intermediate Faithfulness stage. Instead of using an ordinal OT learning algorithm that gradually accepts errors, this section returns to a gradual, GLA-like learner but instead learns a different kind of grammar: a Harmonic Grammar (HG; Legendre et al., 1990a,b; Smolensky and Legendre, 2006; Pater et al., 2007a) Summarizing one key result from Jesney and Tessier (2007, in press), I will first show that this GLA-like learner that uses HG evaluation will in fact reach IF stages, even without a bias for Specific \gg General faith. However this learner is still very sensitive to frequency of violation, because it uses gradual re-ranking. Thus, section 6 will demonstrate how this learner is still susceptible to frequency-induced dangers, because it cannot *revise* its learning from previous errors.

5.1. Weighted constraints as an alternative theory of grammar

The theory of Harmonic Grammar, like Optimality Theory, is a hypothesis about how grammars use constraints to assess linguistic forms. The Harmonic Grammar view of constraint interaction (laid out in Prince and Smolensky, 2004:236 as an alternative to OT) is that each constraint has a numerical weight, and the harmony of each output candidate is the sum of its constraint violations each multiplied by their weights. Following Smolensky and Legendre (2006) and Pater et al. (2007a), I will use an HG grammar in which constraint violations are negative numbers, so that EVAL returns as optimal the candidate with the highest harmony, meaning the number closest to zero (see also Jesney and Tessier, 2007; Keller, 2006; Prince, 2002a).

To see how constraint ranking and weighting differ, consider the example in (44) below. This normal two-by-two OT tableau can also be interpreted in a constraint weighting grammar if each constraint is given a weight (in the top row). The final column shows the ‘score’ for each output—each of its violations multiplied by its constraint weight:

²⁸ One other application of these re-runs is to so-called U-shaped development, where child grammars sometimes regress to what appears to be a less marked grammar after some time of having a more advanced one (see Stemberger et al., 2001; Bleile and Tomblin, 1991; Macken and Ferguson, 1983; Menn, 1976, 1983). Under this approach, regressions could be brought about by running the current Support through the MBCD algorithm and discovering that some rankings are not justified by any stored errors, prompting a return to a more restrictive grammar.

²⁹ I am grateful to Joe Pater, Karen Jesney and Marcin Morzycki for insightful and challenging discussion of the issues in this section, and to an anonymous reviewer for careful comments and suggestions.

44) Weighting ≠ Ranking

Weights:	2	1	Score (sum of violations * weights)
Input	CON1	CON2	
<i>OT winner</i> ↵	Output 1		0(2) + -3(1) = -3
	Output 2	*	-1(2) + 0(3) = -2

↵ *HG winner*

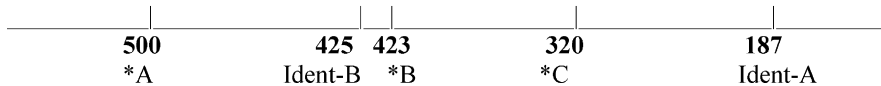
In the familiar OT ranking, Output2's violation of CON1 is fatal, and so Output1 is the winner. But when considered by a HG ranking, Output1's three violation of low-weighted CON2 is enough to rule it out and choose Output 2 as the winner. Following Pater et al. (2007a,b), I interpret constraint violations as negative integers.³⁰

The present paper cannot hope to adequately discuss the range of differences between an ordinal OT system and a weighted Harmonic Grammar, or their consequences: for some of the additional recent literature on the topic (see Coetzee and Pater, in press; Flemming, 2001; Jesney, 2006; Keller, 2006; Tesar, 2007). Whatever the final consensus about their relative merits, further investigation of weighted constraint grammars seems warranted, and has already been fruitful to the investigation of natural language typology, acquisition and implementation. The result relevant to this topic, reported in Jesney and Tessier (2007, in press) is summarized below: that a weighted constraint grammar and a GLA-like learner, will naturally pass through Intermediate Faith stage even without a Specific ≫ General faithfulness bias.

5.2. Deriving the French IF stage using weighted constraints and GLA-like learning

In this view of grammar, every language is characterized by the *weights* on each of its constraints. We can therefore imagine these weights on a number line, just like the ranking values of the GLA's stochastic OT:

45) repeated from 17)



The diagram in (45) can be interpreted as either a stochastic OT grammar or a weighted constraint grammar: the difference will only be in how EVAL uses these numbers to choose optimal input–output mappings. And since a weighted constraint grammar looks like (45), it can also be learned using the same core of the GLA technique already seen. Recall how the GLA learns: given an ERC row with its Ws, Ls and es, the GLA promotes every constraint that prefers the winner and demotes every constraint that prefers the loser.

Thus, we can feed an initial set of constraint values and an initial Support (set of errors) to a GLA-like learner and ask it to gradually increase and decrease these values to fit those errors. What is different from previous sections is how new errors and grammars are evaluated: to choose the loser for any winner, the constraint 'values' from the number line in (54) must be interpreted as weights rather than rankings. This EVAL task can be simulated using the LinearOT evaluation method in recent versions of Praat (Boersma and Weenink, 2007; this method is described in the Praat manual as an implementation of Keller (2006)'s LinearOT).³¹ In this mode, the learner moves its constraints values up and down in response to errors, and the effects of new constraint values are evaluated using a constraint weighting grammar. This will be illustrated at each crucial stage of learning, beginning with the initial state.

In this simulation, markedness constraints all begin with an initial weight of 100, and all faithfulness constraints with an initial weight of 0.³² To make the example as simple as possible, I fed the learner only the three crucial constraints from the French example: *COMPLEXONSET, MAX-IO and MAX-IO/STRESSED. The two errors I gave the learner were *gros* and *brisé*, which provide complex onsets in a stressed and unstressed syllable, respectively.

³⁰ Notice that constraint order in a HG tableau is not as visually instructive as in OT: unlike the candidate 'winnowing' that can be applied left-to-right when reading an OT tableau, every constraint's violations in an HG system may need to be considered to find the optimal output.

³¹ As noted by the term *GLA-like* used throughout, this LinearOT learner has a slightly different update rule from the OT-GLA method. Despite Praat's use of the Keller term, I have chosen not to use the term 'LinearOT learning' here, because I use the labels OT and HG to refer to grammars that use ranked constraints with strict domination on the one hand and weighted constraints on the other; for present purposes this seems a crucial distinction to keep in the terminology.

³² For why faith should start as low as zero in this model, see Jesney and Tessier (in press).

46) *The two errors*

<i>target</i>	<i>winner ~ loser</i>	*COMPLEX	MAX/STRESSED	MAX
<i>gros</i>	gʁo ~ ko	L	W	W
<i>brisé</i>	bʁi'ze ~ pi'ze:	L	e	W

Plasticity was set throughout the simulation at 0.1 and both errors were presented to the learner equally often. Here is how the grammar in Praat looked initially:

47) *The initial grammar*

a. Constraint weights:	*COMPLEXONSET	100.00
	MAX	0.00
	MAX/STRESSED	0.00

b. The mappings:

/!gʁo/	*COMPLEX	MAX	MAX/STRESSED
	100	0	0
'gʁo	*		
☞ 'ko		*	*

c. The constraint weight sums:

$$-1*(100) = -100$$

$$-1*(0) + -1*(0) = \mathbf{0}$$

/bʁi'ze/	*COMPLEX	MAX	MAX/STRESSED
	100	0	0
bʁi'ze	*		
☞ pi'ze:		*	

$$-1*(100) = -100$$

$$-1*(0) = \mathbf{0}$$

When this mode of learning is applied to these constraint weights and errors, the weights change just as the ranking values did in section 3, in direct proportion to their violation frequencies. General MAX prefers the winner in both errors, so its weight is increased every time an error is made; MAX/STRESSED prefers the winner only in the second error, so its weight will be increased half as often. Thus, MAX rises twice as fast as its specific counterpart.

After between 400 and 500 trials of learning on this grammar, the ranking values have changed sufficiently to choose a different set of optimal outputs. A typical state of the grammar at this point is illustrated below, after 500 trials—and this grammar represents the French IF stage.

48) *The grammar after 500 trials*

a. Constraint weights:	*COMPLEX	54.778
	MAX	45.222
	MAX/STRESSED	19.582

b. The mappings:

/!gʁo/	*COMPLEX	MAX	MAX/STRESSED
	54.778	45.222	19.582
☞ 'gʁo	*		
'ko		*	*

c. The constraint weight sums:

$$-1*(54.778) = \mathbf{-54.778}$$

$$-1*(45.222) + -1*(19.582) = -64.804$$

/bʁi'ze/	*COMPLEX	MAX	MAX/STRESSED
	54.778	45.222	19.582
bʁi'ze	*		
☞ pi'ze:		*	

$$-1*(54.778) = -54.778$$

$$-1*(45.222) = \mathbf{-45.222}$$

This new grammar in (48) retains onset clusters in a *stressed* syllable only. Looking at the constraint weight sums, we see that this change in optimal output comes from the collective strength of specific and general MAX: while each is less powerful than markedness on its own, their combined strength is enough to 'gang up' on

*COMPLEX, and so prevent deletion in *[ko]. This ‘gang’ effect is precisely what OT strict domination prohibits in constraint interactions.

Since this grammar is still making errors (on ‘*brisé*’), it will continue to learn, i.e. adjust its constraint weights. After about 650 or so trials, the learner has found a stable grammar that does not produce any further errors, illustrated below:

49) *The grammar after 650 trials*

a. Constraint weights:	MAX	51.986
	*COMPLEX:	48.014
	MAX/STRESSED	19.582

b. The mappings:

/gʁo/	MAX 51.986	*COMPLEX 48.014	MAX/STRESSED 19.582
gʁo		*	
'ko	*		*

c) The constraint weight sums:

$$-1*(48.014) = \mathbf{-48.014}$$

$$-1*(51.986) + -1*(19.582) = -71.568$$

/bʁi'ze/	MAX 51.986	*COMPLEX 48.014	MAX/STRESSED 19.582
bʁi'ze		*	
pi'ze:	*		

$$-1*(48.014) = \mathbf{-48.014}$$

$$-1*(51.986) = \mathbf{-51.986}$$

Compared to the IF stage, this final grammar has increased the weight of general MAX so that it is greater than that of *COMPLEXONSET, while the value of MAX/STRESSED has stayed exactly the same. Now that MAX’s constraint weight is sufficiently large, even a single deletion is now enough to outweigh *COMPLEXONSET, so the cluster in the unstressed syllable of ‘*brisé*’ will surface faithfully. This is therefore the end-state grammar of French.

5.3. How the GLA-like learner of Harmonic Grammars reaches IF stages

Given that the learner maps directly between violation frequencies and constraint movement, why was it able to produce the IF stage in the above simulation? The answer comes from the nature of additive constraint interaction: in a sense, it is because every constraint violation of a candidate contributes towards its harmony. In an OT grammar learned by this GLA-like learner, general and specific constraints rise in value until general MAX overcomes markedness, when both errors in, e.g. (46) are resolved and the target state is achieved. But as the weights of general and specific MAX rise in the HG system, there comes a time when general MAX is still not important enough to outweigh markedness, but both general and specific MAX are strong enough to gang up on markedness, and this results in the IF stage.

50) *Necessary weighting inequalities for an IF stage in a Harmonic Grammar*

$$w(\text{Markedness}) > w(\text{GeneralFaith})$$

$$w(\text{SpecificFaith}) + w(\text{GeneralFaith}) > w(\text{Markedness})$$

A further result is that this learner reaches the IF stage *without a ranking bias between faithfulness constraints*; its only initial weighting bias is that markedness outweighs IO-faithfulness.³³ The subset/superset relationship between MAX/STRESSED and MAX is not something the learner must know to find the IF stage—instead, the fact that their violation profiles stand in this relationship is enough for the stage to emerge.

Two points should be emphasized before moving on from this result. The first is that, in contrast to the Error-Selective Learner already seen, this gradual HG learner is *required* to pass through each IF stage along the way to a target language in which a marked structure is allowed in all contexts.³⁴ As mentioned already, the empirical evidence for ubiquitous IF stages is not at all clear, but perhaps little should be made of this result until better data has been acquired.

A different point is that this HG learner’s relies crucially on the GLA style of learning to reach an Intermediate Faith stage. A logical alternative, which will be discussed in section 7.2, is a Harmonic Grammar learner used its ERC rows

³³ And that IO-faith begins at zero.

³⁴ Unless a language had a marked structure in only the *complement* of the privileged context—e.g. that all complex onsets were found *only* in unstressed syllables.

in the spirit of Recursive Constraint Demotion, using them to learn weighting inequalities such as ‘the weight of MAX must be greater than that of *COMPLEXONSET’. If so, the HG learner would be as non-gradual as ever, and one error like ‘*brisé*’ would be enough to skip over the IF stage just as in section 3.

6. Frequency of violation and the problem of Winner Misparses

As has been emphasized, the HG learner of section 5 and ESL of section 4 differ *how* they are gradual. What any GLA-style learner does gradually is change its constraint values: their relative distance from each other represents the core of the learner’s grammatical knowledge. From the ESL viewpoint the situation is almost opposite: what are learned gradually are the errors, as stored both in the Cache and Support, and the current grammar represents nothing but a particular Support-to-ranking mapping done by MBCD. While the ESL learner uses its memory of previous errors to create ever newer grammars, a GLA-like learner has no such memory. The errors that brought it to the present state are neither stored nor even directly recoverable from that ranking.

The GLA-style learner must therefore trust that the frequency with which it gets its errors and their violation profiles will never lead it too far astray, that it will never need a memory for how it reached the current grammar. What I will argue in this section is that one class of learning situations makes such a memory crucial. In these cases, learning from violation frequencies without such a memory can bring the learner not to the wrong *intermediate* stage, but rather the wrong *end-state* grammar. This argument is germane to the goals of this paper because it suggests a fundamental problem with learning via gradual, numerical re-ranking and violation frequencies. While this problem exists for any GLA-style learner, whether using OT or Harmonic Grammar, the two case studies in this section will use HG tableaux to keep up the comparison with section 5.

The general outline of the problem is this. Learners often have to make assumptions about hidden structural information, such as syllable or foot structure, before they can assess an error’s Ws and Ls. If the current grammar makes the wrong assumptions about structure, the resulting ERC rows can include what I will call ‘winner misparses’, and these misparses can lead the learner to a superset language. This kind of misleading evidence from frequency of violation is again the result of constraints whose contexts are in subset/superset relationships—and these can occur even when two *markedness* constraints are at issue. The two examples provided in the next two sections are hypothetical, but given the fairly standard constraints assumed in each case, the claim is that superset languages *could* be found in children’s early phonologies if the right target grammar intersected with some particular statistical properties of a child’s early lexicon. However, the more general point made in this limited space is that such superset languages are certainly possible, if not probable, and thus are a challenge which any learner should be prepared to face.

Winner misparses can happen to any of the learners discussed here; the question is how a learner can recover from them, at the point when a winner misparse has been resolved. These examples show that even a cautious learner cannot take its errors at face value, and therefore that a successful gradual learner must still store errors and be able to reason about them later. From the present perspective, this is something that the ESL alone is well-equipped to do.

6.1. Winner misparses and relations among Faith constraints violations³⁵

Suppose that the learner is acquiring a language with coda devoicing: so that onset segments can be voiced or voiceless, while codas can only be voiceless. In the previous sections of this paper, such a set of weights would have been characterized as an IF stage,³⁶ but it also characterizes the end-state of learning for languages like Dutch in which coda voicing is neutralized. As we saw in (50), an IF pattern requires two weighting inequalities. To translate to the contextual neutralization of voicing rather than complex onsets, we can plug the constraints of Lombardi (1999) on voicing into the inequalities, as in (51):

51) A target grammar – contextual neutralization of voicing in coda obstruents

$$\begin{aligned} & \#* \text{VOICEDOBSTRUENT} > \# \text{IDENT}[\text{VOICE}] \\ & \#(\text{IDENT}[\text{VOICE}]-\text{ONSET}) + \# \text{IDENT}[\text{VOICE}] > * \text{VOICEDOBSTRUENT} \end{aligned}$$

³⁵ I am grateful to John McCarthy for suggesting this example to me.

³⁶ Such an intermediate stage of obstruent devoicing is reported for the English-learning child in Smith (1973).

Assuming the $M \gg F$ initial state, the learner of this language will initially devoice voiced obstruents under the pressure of *VOICED_{OBS} so that a target form like [ˈkɪbla] will come out unfaithfully as [ˈkɪpla]. Under one other assumption we will see in a moment, this means that they will create ERC rows like (52):

52) *Learning onset voicing: the right ERC row*

winner ~ loser	*VOICED _{OBS}	IDENT[VOICE]-ONS	IDENT[VOICE]
[ˈkɪ. bla] ~ [ˈkɪ. pla]	L	W	W

The crucial assumption encoded in (52) is that the voiced target segment that gets devoiced is syllabified as an onset. This assumption comes from other rankings in the grammar that determine the optimality of each syllabification: for example, the relative ranking of STRESS-TO-WEIGHT (which is violated by a stressed light syllable like [ˈkɪ] and NoCODA. As the two tableaux below show, getting the right ERC row in (52) requires that w_{NoCODA} is greater than $w_{\text{STRESS-TO-WEIGHT}}$ (numerical weights here chosen arbitrarily)³⁷:

53) *Two different syllabifications of the ERC forms*

/ˈkɪbla/	STRESS-TO-WEIGHT 110	NoCODA 90		/ˈkɪbla/	NoCODA 110	STRESS-TO-WEIGHT 90	
[ˈkɪ. bla]	*!		= -110	☞ [ˈkɪ. bla]		*	= -90
☞ [ˈkɪ b .la]		*	= -90	[ˈkɪ b .la]	*!		= -110

But what if our learner currently has a greater weight assigned to STRESS-TO-WEIGHT rather than NoCODA? This will cause them to represent this same error on ‘kɪbla’ differently, not as in (52) but as in (54):

54) *Learning onset voicing with the wrong ERC row*

winner ~ loser	*VOICED _{OBS}	IDENT[VOICE]-ONS	IDENT[VOICE]
[ˈkɪ b .la ~ ˈkɪ p .la]	L	e	W

Comparing the right and wrong ERC rows above, the important difference is already apparent: IDENT(VOICE)-ONSET does not assign a W in (54), because the segment which is unfaithful to voicing is not an onset. So even though the HG learner does not need any biases to get to the IF stage, the error in (54) will still cause problems—because the learner *does not realize that the correct end-state grammar resolves this error*.

An illustration of this problem begins with an early ranking, at which markedness is above faithfulness and general faithfulness has already climbed somewhat above specific faith. Errors like (54) will cause the demotion of *VOICED_{OBS} and the promotion of general IDENT[VCE].³⁸

55) *The re-ranking effect of the ERC row, at the early stage:*

\rightarrow		\leftarrow
100 98	22 20	10
*VOICED _{OBS}	ID[VCE]	ID[VCE]-ONS

When armed with errors like those in (52), we have already seen that the HG-GLA will reach the IF stage, in which both faith constraints can gang up on *VOICED_{OBS}, but general IDENT-VOICE cannot do it alone. For a language

³⁷ I am leaving aside some technical discussion about how the current grammar assigns hidden structure to winners: see esp. [Tesar and Smolensky \(2000\)](#). What is crucial here is that the same current grammar syllabifies both the winner form and the loser, so that within an ERC row, syllabification will remain constant.

³⁸ Although the error in (55) does not cause the HG-GLA learner to change the weighting of IDENT[VCE]-ONSET, I have given it a value greater than the initial zero weight. This is because it seems safe to assume that the language will provide learners with forms other than (54) in which voiced obstruents do occur in an unambiguous onset position (e.g. [ˈbɪk.la] or [ɪ.ˈba]), causing errors in which ID-ONSET does indeed prefer the winner and so has its weight increased.

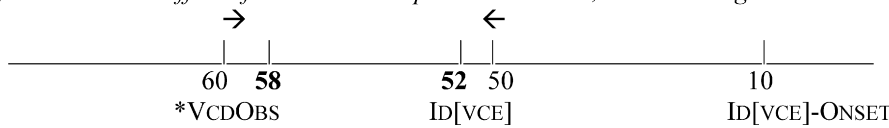
like Dutch where only onset obstruents can be voiced, an IF grammar in fact characterizes the target—and yet this current grammar’s beliefs about syllabification mean that the learner is still making errors:

56) *The continuing error at the IF stage*

/kɪbla/	*VOICEDOBS 60	ID(VOICE) 50	ID(VOICE)-ONSET 10	
winner: [kɪb.la]	*!			= -60
loser: [kɪp.la]		*		= -50

Even though the learner has reached the target ranking, it has not resolved this error because of its mistaken syllabification. And so learning continues, *VOICEDOBSTRUENT continues to be demoted, and general IDENT-VOICE continues to rise:

57) *The continued effect of the winner misparsed ERC row, at the IF stage:*



This learning will continue until one of two things happen: the weight of general IDENT(VOICE) gets above *VOICEDOBSTRUENT, or the learner learns the right weightings of STRESS-TO-WEIGHT and NoCODA. The danger is that if the former happens first, the learner will have reached the wrong final state:

58) *The incorrect final grammar*

$$w_{IDENT-[VOICE]} > w_{*VCD OBS} > w_{IDENT-[VOICE]-ONSET}$$

The pathology here is this first inequality: with $w_{IDENT-VOICE}$ being greater than $w_{*VOICEDOBSTRUENT}$, voiced obstruents will be incorrectly protected outside of onset positions (e.g. /kɪb/ → *[kɪb]).

6.2. *Escaping the superset grammars of winner misparses*

We have seen thus far that a learner, working from misleading errors, can be deluded into choosing a less restrictive grammar than the intended target. This is just as much a problem for the Error-Selective BCD learner from section 4: with a misyllabified error like (54) in the Support, an ordinal learner can just as easily end up with the OT version of the grammar in (58). The issue is whether the learner has a way to *undo* this error.

The Error-Selective Learner that has acquired the incorrect grammar in (58) will only have done so because it has added the error of (54) to its Support—and because (54) fails to reveal that ranking a specific faith constraint above markedness is sufficient to resolve this error. Because the problem is thus localized, it can be fixed—all that must happen is an update to the Support.

The crucial step for the Error-Selective Learner will be adopting the correct language-specific ranking of NoCODA ≫ STRESS-TO-WEIGHT. When this ranking has been changed, the learner now knows something new about the target language that they did not know when they assigned structure to their Support’s winners. Thus, we can equip the ESL learner with a requirement that from time to time, it should re-assign hidden structure to the winners and losers in their ERC rows using the *current* grammar, and then re-assign Ws and Ls to those forms that have changed in any way. In the present case, the new grammar in which NoCODA dominates STRESS TO WEIGHT will resyllabify all intervocalic clusters as complex onsets regardless of their stress, so that [kɪb.la] will be resyllabified as [ki.bla] and the ERC row in (54) will be updated to look like (59) (repeated below from (52):

59) *The correct ERC row*

winner ~ loser	*VOICEDOBS	IDENT[VOICE]-ONS	IDENT[VOICE]
[kɪ.bla] ~ [kɪ.pla]	L	W	W

With all intervocalic voiced obstruents resyllabified in this way, the Support will now be rid of any errors that provide evidence for the OT ranking of IDENT[VOICE] \gg *VOICEDOBSTRUENT. The next time the Error-Selective Learner runs MBCD on this new Support with errors like (59), the familiar ranking bias for installing specific faith constraints rather than general ones will be sufficient to ensure that the learner reaches the correct OT ranking as in (60):

- 60) *Correct ranking learned by MBCD given errors like 59)*
 IDENT[VOICE]-ONS \gg *VOICEDOBS \gg IDENT[VOICE]

Note that the MBCD algorithm will not know how many times it made the old, misleading error from (54), nor could it do anything with that knowledge even if it did. Once the Support has the new syllabification for ['kɪ.blə], this learner has forgotten (54), and the next cycle of re-ranking will immediately result in the correct OT grammar. Notice also that the Error-Selective Learner will return to the more restrictive grammar in (60) even after it has already *stopped making errors* on forms like ['kɪ.blə]. The trigger for adopting the more restrictive grammar is not due to anything about voiced obstruents per se, but rather a consequence of the learner's check on the match between Support and grammar.

In contrast, a GLA-style learner does not have any straightforward way of moving beyond the superset grammar of (58). Suppose that, at some point after (58) have been adopted, the learner gradually switches the ranking of the two structural constraints, so that now NoCODA *does* outweigh STRESS-TO-WEIGHT. Now what should it do about the winner misparse ERC row for 'kibla'? The constraint with the incorrect position is IDENT(VOICE), and faithfulness to voicing has nothing in principle to do with either of the two markedness constraints whose rankings have been reversed, except for their indirect effect on the errors that caused re-ranking. How now should the learner realize that some of its earlier errors should not have prompted the re-ranking of *VOICED OBSTRUENT and IDENT-VOICE? This is a way in which an incremental algorithm like the GLA's disconnect between errors and their re-ranking effects seems to cause real problems. In fact, this learner has no way of demoting below a particular constraint—it only moves constraints along its numerical scale, without reference to the (ordinal) position of any others. And this learner's problem is again related again to violation frequencies, because its learner remembers *how many times* each constraint was shown to need promoting or demoting, rather than *why* it needed to be promoted or demoted or which constraints it needed to outrank. As it stands, it is not clear how a GLA-style learner should incorporate any of the necessary reasoning into its method of gradual re-ranking.³⁹

6.3. Winner misparses and relations among markedness constraint violations

This section presents a different winner misparse trap in which learning straight from frequency of violation can cause problems. While again illustrating the central point of this paper, it also differs in at least two important ways from the previous example. First, it uses a different kind of hidden structure, relying on feet rather than syllables.⁴⁰ As a reviewer points out, mis-syllabifications like those in the previous section will not present a trap for the learner if, following Côté (2000), Jun (1995), Steriade (1999) and others, phonotactic constraints are characterized not with reference to syllable structure but rather by position in the segmental string. Under this view, the constraint IDENT[VOICE]/ONSET will be re-defined as IDENT-[VCE]/PRE-SONORANT (for example), so that the ERC row [kɪblə ~ kɪplə] will be assigned a W by both faithfulness constraints regardless of syllabification. Second, this example triggers a winner misparse using markedness constraints, to demonstrate that it is not faithfulness constraints themselves that are the ultimate cause of this problem.

In this example, the language in question has an allophonic alternation between aspirated and unaspirated voiceless stops, according to which the aspirates appear in foot-initial position and the plain stops appear elsewhere. (These are

³⁹ One possibility, originally suggested to me by Kie Zuraw, might be that the GLA learner could include a persistent pressure to demote *all* IO faithfulness constraints – or perhaps in this case, all *general* faithfulness constraints – at all times. (See also a related suggestion in the OTSoft Manual.) The gist of this proposal would be that over time, any general IO-faithfulness constraints would always be sinking back down the scale of values (or weights). Two crucial issues for any such approach, once fleshed out, would be (a) how a learner would know that it should *stop* trying to demote general faithfulness, in a language where, e.g. IDENT[VOICE] really does dominate *VOICED OBSTRUENT, rather than forever cycling through error generation and resolution, and (b) whether the resulting error patterns would match attested developmental stages.

⁴⁰ Thanks to Joe Pater for suggesting this particular example to me.

nearly the facts of English aspiration.⁴¹) To capture this basic allophonic pattern, I will use the two markedness constraints in (61) below:

- 61) a) *UNASPIRATED VOICELESS STOP/FOOT-INITIAL (shortened to ASP/FT-INIT)
 b) *ASPIRATE (shortened to *ASP)

The constraint in (61a) is presumably a positional strengthening constraint (see esp. Smith, 2002), and the one in (61b) is a context-free markedness constraint, necessary to derive the lack of aspirates in many languages. In a classic OT grammar, this kind of allophony comes from a ranking of specific markedness \gg general markedness \gg faithfulness. As it turns out, this ranking is simply mirrored in the HG system of constraint weights:

- 62) *Necessary HG and OT grammars for allophonic aspiration*
 a) OT: ASP/FT-INITIAL \gg *ASP \gg IDENT-ASP
 b) HG: w ASP/FT-INITIAL $>$ w *ASP $>$ w IDENT-ASP

As in previous sections, we begin with the initial state of both markedness constraints weighted equally high and IDENT-ASP low (i.e. at zero). With this initial state, the learner can reach this grammar in (62) by making de-aspiration errors, like the one in (63) below. This error will come about if the more general markedness constraint (*ASP) has a weight slightly greater than the specific one (ASP/FT-INITIAL), so that aspiration is removed from a foot-initial segment:

63) *Learning the distribution of voicing: the right ERC row*

winner ~ loser	*ASPIRATE	ASP/FT-INITIAL	IDENT(ASP)
	101	99	0
[bo(t ^h ego)] ~ [bo(tego)]	L	W	W

From this ERC row, the learner will increase the weight of ASP/FT-INITIAL and decrease the weight of *ASPIRATE, and so the right ranking will quickly be reached. But as with the previous example of syllabification: this ERC row is only as right as the learner's current belief about foot structure. The output form in (63) above, *bo^hégo*, was chosen precisely because it contains a classic structural ambiguity—is this medial stressed syllable the result of an initial iambic foot, or a final trochaic one? The choice between these two analyses will be made by the relative weights of footing constraints, i.e. TROCHEE vs. IAMB. If currently the learner's grammar includes the incorrect weighting w IAMB $>$ w TROCHEE, then the learner will have constructed the following incorrect ERC row⁴²:

64) *The wrong assignment of foot structure, and the wrong ERC row*

winner ~ loser	*ASPIRATE	ASP/FT-INITIAL	IDENT(ASP)
	101	99	0
[(bo(t ^h e)go)] ~ [(bo(te)go)]	L	e	W

What will be the result of this winner misparse in (64) above? As a result of this error, the HG-GLA learner will increase the weight of *ASP and decrease the weight of IDENT(ASP) only—it will leave the value of ASP/FT-INITIAL unchanged, since in this learner's eyes the aspirated stop of *bo^hégo* is not foot-initial. If the learner's incorrect assumption about the foot structure of medially stressed three-syllable words persists long enough, errors of this sort will prompt the learner to invert the relative weightings of IDENT(ASP) and *ASP, resulting in an end-state grammar as in (65):

⁴¹ The complications raised by aspiration in, e.g. word-initial position as in [p^ho]tato, and in morphologically complex contexts, will not be dealt with here: see, e.g. Jensen (1993) and Davis (2003).

⁴² A reviewer points out that using this constraint Asp/Ft-Initial allows for the possibility of the error in (64) being the *right* parse in some language—that is, a language which builds iambs, but which requires aspiration among just those unstressed stops which are foot-initial. In the absence of any such typological evidence, it may be that this contextual markedness constraint is not the right one for the task.

- 65) *Eventual result of HG-GLA learning from the wrong ERC row in 64)*
 $w_{\text{ASP/FT-INITIAL}} > w_{\text{IDENT(ASP)}} > w_{*\text{ASPIRATE}}$

Compared to the target language in (62), this grammar is in fact a superset language. Rather than allophonic aspiration, this grammar creates positional neutralization of aspiration: at the beginning of a foot, voiceless obstruents are correctly required to be aspirated, but anywhere else they incorrectly show *contrastive* aspiration:

66) *Lack of restrictiveness in the grammar of 65)*

/bo'dek ^h o /	ASP/FT-INITIAL	IDENT(ASP)	*ASPIRATE	
	99	60	40	
☛ bo'dek ^h o			*	= -40
bo'dego		*!		= -60

As with the previous example, the Error-Selective Learner will retain this superset grammar only as long as it has the ERC row of (64). Once the learner establishes that Trochee outranks Iamb, their re-parsing of winners like [bo^hego] will assign the correct foot structure as in (63). The next time the MBCD algorithm is fed the revised Support, its first bias for installing markedness constraints whenever they do not prefer Losers will now install ASP/FT-INITIAL in the first stratum, resolving the error right away. With all cases of aspiration resolved, the ranking bias will now be able to install *ASPIRATE above IDENT(ASP), yielding the correct grammar in (62a). For the GLA, however, there is again no clear way in which reversing the values (or weights) of IAMB and TROCHEE should or could cause any re-ranking of constraints on aspiration.

7. Concluding discussion

7.1. Summarizing the results across learning alternatives

This paper represents one attempt to combine the results of natural language acquisition and learnability theories. It has demonstrated that a class of attested intermediate stages in phonological development cannot be produced by an OT learner that re-ranks its constraints in direct proportion to the frequency of constraint violations and satisfaction in its errors. More generally, this paper has brought into focus some issues facing a constraint-based learner which aims to be gradual, restrictive and sensitive to frequency.

While the paper began with the most basic OT-GLA algorithm, its eventual result has been to propose a very different method of constraint-based gradual learning: Error-Selective Learning using Multiply-Biased Constraint Demotion. section 2 introduced the evidence from L1 developmental stages; these cases suggest that learners must be able to innovate grammars that are *more* restrictive than their targets, in a way that contradicts the frequency with which constraints are violated. The resulting claim has been that the Error-Selective Learner is better equipped to get from available errors to restrictive rankings, predominantly because its sensitivity to violation frequency is mediated through its ranking biases. In addition, section 6 introduced evidence from potential end-state superset grammars to suggest that no matter how restrictive a gradual re-ranking learner is, it must still be able to undo previous rankings in a *non-gradual* way. These examples were used to argue that the Error-Selective Learner's reliance on the Support's stored errors is justified, and in fact necessary.

One result in the literature that is related to the present discussion is that of Pater (*in press*), which reports a certain clusters of ERC rows for which the OT-GLA fails to converge on a grammar. The diagnosis that Pater provides is a difficulty in handling the Credit Problem (Dresher, 1999), i.e. to determine which W-preferring constraints need to be ranked above which L-preferring ones. On the one hand, the dilemma of winner misparses is certainly related to the Credit Problem—because one thing a learner like the GLA cannot do is reverse its poor placement of credit for winner ~ loser pairs *after it has made errors and forgotten them*. Pater (*in press*) demonstrates that a GLA-like learner that uses constraint weighting *does* converge on a grammar for this data. On the other hand, section 6 demonstrated that winner misparses cause difficulties for the GLA's method of gradual learning, not for a particular method of grammatical evaluation (constraint ranking vs. weighting).

7.2. A final learning alternative

As a reviewer rightly points out, this paper has not yet discussed a fourth logical possibility in the search for both a restrictive grammar and a gradual learner. The approach supported here is a combination of an OT grammar and a non-gradual learning algorithm, namely MBCD. The two alternatives to which it has been compared both involved gradual GLA-like learning, combined with grammars that either rank or weight constraints, using OT or HG respectively. The remaining combination is a learner that builds Harmonic Grammars and yet does *not* use a gradual algorithm, but rather one that resolves its errors as MBCD does. One might well ask whether an Error-Selective HG learner, using a non-gradual algorithm to re-weight constraints, could achieve the same results as the ESL of section 4—or even better results.

To my knowledge, the only learner on the market which resolves errors to build a Harmonic Grammar is found in Pater et al. (2007a). They show how a HG can be interpreted as a linear system, so that the problem of grammar learning can be solved using the simplex algorithm, a highly efficient optimization tool from applied mathematics. Their initial method, implemented in the online software tool HaLP (Potts et al., 2007b), simply learns a Harmonic Grammar that resolves a set of errors (if possible) while minimizing the total constraint weights. Furthermore, a newer version of HaLP can impose weighting biases on constraints,⁴³ such as a bias to keep faith weighted as low as possible (Michael Becker, p.c.)—thereby making the HaLP learner restrictive as well.

However, it is in fact not possible to simply re-imagine Error-Selective Learning using HG by replacing the MBCD algorithm with the HaLP method. The reason is that the Error-Selective Learning approach proposed in section 4 is in part tied to the workings of MBCD itself, which in turn are tied to properties of Optimality Theory and the strict ranking of constraints. Step Two of ESL, spelled out in its final form in (40) of section 4.2.3, applies MBCD to a set of potential ERC rows until one has been resolved, and then discards all the others. Recall again the two errors that the learner had to choose between in (39): *gros*, with a complex onset cluster protected by both MAX/STRESSED and MAX, and *gruau*, with a complex onset protected only by general MAX. In this relevant case, the ESL procedure allows the specific \gg general IO-faith bias of MBCD to choose which potential ERC row is resolved first—this was spelled out in the reasoning that gets the French learner from tables (39) to (40), and chooses *gros*.

How could this procedure be mimicked by a non-gradual HG learner, attempting to choose between potential ERC rows? Given a HaLP-style algorithm which preferred to keep faithfulness constraints weighted low, an error like *gros* would indeed prompt the learner to successfully build a grammar with the weighting conditions from section 5's example (50): the weights of specific and general faith constraints would be increased sufficiently for them to gang up on *COMPLEXONSET and protect onset clusters only in stressed contexts. But how could this ESL's Step Two be revised to choose *gros* as the right ERC row to learn from, rather than *gruau*? In the OT-ESL approach, the decision is made by both the MBCD's bias for specific \gg general faith, and the installation of MAX/STRESSED into an OT ranking, whose presence in a stratum n ensures that no subsequent constraints installed in stratum $n + 1$ or further down will ever *unresolve* the error. Even if the HaLP learner was further biased to keep the weights of general faith constraints even lower than specific ones, mimicking the second MBCD bias from section 4, the learner would still need a way to choose between *gros* and *gruau* using that bias. And since a Harmonic Grammar can only choose between two candidates when *every* constraint's weighted contribution has been assessed, this learner would need to determine that the sum of every constraint's weight will be lower if we choose *gros* rather than *gruau*. Choosing the right Potential Best Errors would mean determining the total constraint weights created by adding each PBE to the existing Support in turn, and finally choosing to retain just those ERC rows that result in the smallest total weights.

In this light, the HG-ESL approach does not seem like a promising alternative to the OT version proposed in section 4. Unlike the GLA-style HG learner of section 5, the Error-Selective HG learner will still require a specific $>$ general faith bias, and it additionally requires the calculation of multiple *entire potential grammars* on every cycle of learning. However, in the spirit of this paper's broader goals, I consider one interesting avenue of future research to include possible alternative methods of online HG learning, using something like Error-Selective Learning, and their potential benefits and consequences.

⁴³ Within the linear programming method, one method of biasing constraints to be low-weighted is to increase their coefficients in the objective function (Christopher Potts, p.c.); on this objective function, see Pater et al. (2007a) and references therein.

7.3. Current conclusions

The largest goal of this paper is to reveal that frequencies are not always reliable information to learn from—and that frequency information needs to be mediated by a grammatically informed learner to correctly steer (and sometimes re-evaluate) development. In addition, this work raises a set of new questions about the kinds of empirical data needed to distinguish between learning proposals in the natural language context. Finding some of these answers may indeed prove difficult, but they represent the ultimate test for this research program.

Acknowledgements

Thanks to Michael Becker, Karen Jesney, Marnie Krauss, John McCarthy, Marcin Morzycki, Mits Ota, Johanne Paradis, Joe Pater, Tamara Sorenson Duncan, Shelley Velleman and participants at BUCLD30 and WCCFL25, as well as Hildibrandt Barca and Tracy O'Brien. Special thanks to two anonymous reviewers for important challenges and alternatives that have greatly improved this material. As always, none of the above are responsible for any of my errors.

References

- Angluin, D., 1980. Inductive inference of formal languages from positive data. *Information and Control* 45, 117–135.
- Apoussidou, D., 2007. The learnability of metrical phonology. LOT-dissertation Series, #148. http://www.fon.hum.uva.nl/diana/The_Learnability_of_Metrical_Phonology.pdf.
- Beckman, J.N., 1998. Positional faithfulness. Doctoral Dissertation. University of Massachusetts Amherst.
- Berwick, R., 1985. *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA.
- Bleile, K., Tomblin, J.B., 1991. Regressions in the phonological development of two children. *Journal of Psycholinguistic Research* 20 (6), 483–499.
- Boersma, P., 1997. Functional phonology: formalizing the interactions between articulatory and perceptual drives. Ph.D. Dissertation. University of Amsterdam/Holland Academic Graphics, The Hague.
- Boersma, P., Levelt, C., 2000. Gradual constraint-ranking learning algorithm predicts acquisition order. In: *Proceedings of 30th Child Language Research Forum*. Stanford University, Stanford: CSLI.
- Boersma, P., Apoussidou, D., 2004. Comparing different optimality-theoretic learning algorithms for Latin stress. In: Schmeiser, B., Chand, V., Kelleher, A., Rodriguez, A. (Eds.), *Proceedings of the WCCFL23*. Cascadilla Press, Somerville, MA, pp. 29–42.
- Boersma, P., Hayes, B., 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32, 45–86.
- Boersma, P., Weenink, D., 2007. Praat: Doing Phonetics by Computer (Version 4.5.18) [Computer Program].
- Booij, G., 1995. *The Phonology of Dutch*. Clarendon Press, Oxford.
- Coetzee, A., Pater, J., in press. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory*.
- Côté, M.-H., 2000. Consonant cluster phonotactics: a perception-based approach. MIT Linguistics Dissertations. MIT Working Papers in Linguistics, Cambridge, 367 pp.
- Curtin, S., 2002. Representational richness in phonological development. Ph.D. Dissertation. University of Southern California.
- Curtin, S., Zuraw, K., 2001. Explaining constraint demotion in a developing system. In: Skarabela, B., Fish, S., Do, A.H.-J. (Eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development*. Cascadilla, Somerville, MA.
- Davis, S., 2003. Capitalistic vs. militaristic: the paradigm uniformity effect reconsidered. In: Downing, L., Hall, T.A., Raffelsiefen, R. (Eds.), *Paradigms in Phonological Theory*. Oxford University Press, Oxford, pp. 107–121.
- Demuth, K., 1996. Stages in the acquisition of prosodic structure. In: Clark, E. (Ed.), *Proceedings of the 27th Child Language Research Forum*. Stanford University: CSLI, pp. 39–48.
- Dresher, B.E., Kaye, J., 1990. A computational learning model for metrical phonology. *Cognition* 34.2, 137–195.
- Dresher, B.E., 1999. Charting the learning path: cues to parameter setting. *Linguistic Inquiry* 30.2, 27–67.
- Fikkert, P., 1994a. On the acquisition of prosodic structure. Ph.D. Dissertation. University of Leiden.
- Echols, C.H., Newport, E., 1992. The role of stress and position in determining first words. *Language Acquisition* 2, 189–220.
- Fikkert, P., 1994b. On the acquisition of prosodic structure. Ph.D. Dissertation. HIL Dissertations 6. Leiden University/Holland Academic Graphics, The Hague.
- Flemming, E., 2001. Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18 (1), 7–44.
- Gerken, L.A., 1996. Prosodic structure in young children's language production. *Language* 72, 683–712.
- Gibson, E., Wexler, K., 1994. Triggers. *Linguistic Inquiry* 25.3, 407–454.
- Gnanadesikan, A., 1995/2004. Markedness and faithfulness constraints in child phonology. In: Kager, R., Pater, J., Zonneveld, W. (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press, Cambridge, UK.
- Goad, H., Rose, Y., 2004. Input elaboration, head faithfulness and evidence for representation in the acquisition of left-edge clusters in West Germanic. In: Kager, R., Pater, J., Zonneveld, W. (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press, Cambridge, UK.
- Goldwater, S., Johnson, M., 2003. Learning OT constraint rankings using a maximum entropy model. In: *Proceedings of the Workshop on Variation within Optimality Theory*. Stockholm University.

- Gouskova, M., 2003. Deriving economy: syncope in optimality theory. Ph.D. Dissertation. University of Massachusetts Amherst. GLSA Publication, Amherst, MA.
- Hayes, B., 2004. Phonological acquisition in optimality theory: the early stages. In: Kager, R., Pater, J., Zonneveld, W. (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press, Cambridge, UK.
- Hayes, B., Londe, Z.C., 2006. Stochastic phonological knowledge: the case of Hungarian Vowel Harmony. *Phonology* 23, 59–104.
- Ito, J., Mester, A., 1999. The structure of the phonological lexicon. In: Tsujimura, N. (Ed.), *The Handbook of Japanese Linguistics*. Blackwell, Oxford, pp. 62–100.
- Jaeger, G., 2007. Maximum entropy models and stochastic optimality theory. In: Zaenen, A., Simpson, J., King, T.H., Grimshaw, J., Maling J., Manning, C. (Eds.), *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. CSLI Publications, Stanford.
- Jarosz, G., 2006. Rich lexicons and restrictive grammars—maximum likelihood learning in optimality theory. Ph.D. Dissertation. Johns Hopkins University, Baltimore, MD.
- Jensen, J., 1993. *English Phonology*. John Benjamins Publishing Company.
- Jesney, K., 2006. Emergent OO-faith stages in a weighted constraint system. Ms. University of Massachusetts Amherst.
- Jesney, K., 2007. The locus of variation in weighted constraint grammars. Poster presented at the Workshop on Variation, Gradience and Frequency in Phonology, Stanford University, Stanford, CA.
- Jesney, K., Tessier, A.-M., 2007. Re-evaluating learning biases in Harmonic Grammar. In: M. Becker (Ed.), *University of Massachusetts Occasional Papers 36: Papers in Theoretical and Computational Phonology*. GLSA, Amherst, MA.
- Jesney, K., Tessier, A.-M., in press. Gradual learning and faithfulness: consequences of ranked vs. weighted constraints. In: Abdurrahman, M., Schardl, A., Walkow, M. (Eds.) *Proceedings of the 38th Meeting of the North East Linguistic Society*.
- Jun, J., 1995. Perceptual and articulatory factors in place assimilation: an optimality theoretic approach. Ph.D. Dissertation. UCLA. UCLA Occasional Papers in Linguistics, Los Angeles, CA.
- Kehoe, M., 2000. Truncation without shape constraints: the latter stages of prosodic acquisition. *Language Acquisition* 8 (1), 23–67.
- Kehoe, M., Stoel-Gammon, C., 1997. Truncation patterns in English-speaking children's word productions. *Journal of Speech, Language, and Hearing Research* 40, 526–541.
- Kehoe, M., Hilaire-Debove, G., 2004. The structure of branching onsets and rising diphthongs: evidence from the acquisition of French. In: Brugos, A., Micciulla, L., Smith, C.E. (Eds.), *Proceedings of the BUCLD 28*. Cascadia Press, Somerville, MA, pp. 282–293.
- Keller, F., 2006. Linear optimality theory as a model of gradience in grammar. In: Fanselow, G., Féry, C., Vogel, R., Schlesewsky, M. (Eds.), *Gradience in Grammar: Generative Perspectives*. Oxford University Press, Oxford.
- Legendre, G., Miyata, Y., Paul, S., 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*. Lawrence Erlbaum, Cambridge, MA, pp. 884–891.
- Legendre, G., Miyata, Y., Smolensky, P., 1990b. Harmonic Grammar—a formal multi-level connectionist theory of linguistic wellformedness: theoretical foundations. In: *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum, Cambridge, MA, pp. 388–395.
- Legendre, G., Sorace, A., Smolensky, P., 2006. The optimality theory—Harmonic Grammar connection. In: Smolensky, P., Legendre, G. (Eds.), *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, vol. I. Cognitive Architecture. MIT Press, Cambridge, MA, pp. 903–966.
- Levelt, C.C., van de Vijver, R., 2004. Syllable types in cross-linguistic and developmental grammars. In: Kager, R., Zonneveld, W., Pater, J. (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press, Cambridge, UK.
- Lombardi, L., 1999. Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory* 17, 1999.
- Macken, M., Ferguson, C., 1983. Cognitive aspects of phonological development: model, evidence and issues. In: Nelson, K. (Ed.), *Children's Language*, vol. 3. Erlbaum, Hillsdale, NJ.
- McCarthy, J.J., 1998. Morpheme structure constraints and paradigm occultation. In: Catherine Gruber, M., Higgins, D., Olson, K., Wysocki, T. (Eds.), *Proceedings of the Chicago Linguistic Society*, 5. vol. II: The Panels. CLS, Chicago.
- McCarthy, J.J., 2003. OT constraints are categorical. *Phonology* 20, 75–138.
- Menn, L., 1976. Pattern, control, and contrast in beginning speech: a case study in the development of word form and word function. Ph.D. Dissertation. University of Illinois, Urbana.
- Menn, L., 1983. Development of articulatory, phonetic, and phonological categories. In: Butterworth, B. (Ed.), *Language Production*. Academic Press, New York.
- Pater, J., 1997. Minimal violation and phonological development. *Language Acquisition* 6 (3), 201–253.
- Pater, J., 2008a. Optimization and linguistic typology. Ms. University of Massachusetts, Amherst. (ROA # 982-0708).
- Pater, J., 2008b. Gradual learning and convergence. *Linguistic Inquiry* 39/2, 334–345.
- Pater, J., Bhatt, R., Potts, C., 2007a. Linguistic Optimization. Ms. University of Massachusetts Amherst (ROA 872-1006).
- Pater, J., Potts, C., Bhatt, R., 2007b. Harmonic grammar with linear programming. Ms. University of Massachusetts Amherst (ROA 872-1006).
- Pearl, L., 2007. Necessary bias in natural language learning. Doctoral Dissertation. University of Maryland.
- Pearl, L., 2008. Putting the emphasis on unambiguous: the feasibility of data filtering for learning English metrical phonology. In: Chan, H., Jacob, H., Kapia, E. (Eds.), *BUCLD 32: Proceedings of the 32nd Annual Boston University Conference on Child Language Development*. Cascadia Press, Somerville, MA, pp. 390–401.
- Potts, C., Becker, M., Bhatt, R., Pater, J., 2007. HaLP: Harmonic Grammar with Linear Programming, Version 2. Software available online at <http://web.linguist.umass.edu/~halp/>.
- Prince, A., 1983. Relating to the grid. *Linguistic Inquiry* 14.1, 19–100.
- Prince, A., 1997. 'Paninian Relations'. Invited Talk. Department of Linguistics, University of Massachusetts Amherst.

- Prince, A., 2002a. Arguing optimality. In: Coetzee, A., Carpenter, A., de Lacy, P. (Eds.), *Papers in Optimality Theory II*. GLSA, Amherst, MA, pp. 269–304.
- Prince, A., 2002b. Entailed ranking arguments. Ms. Rutgers University, New Brunswick, NJ (ROA 500-0202).
- Prince, A., Smolensky, P., 1993/2004. *Optimality theory: constraint interaction in generative grammar*. RuCCS Technical Report 2. Rutgers University Center for Cognitive Science, Rutgers University, Piscataway, NJ (Revised version published by Blackwell).
- Prince, A., Tesar, B., 2004. Learning phonotactic distributions. In: Kager, R., Zonneveld, W., Pater, J. (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press, Cambridge, UK.
- Pulleyblank, D., Turkel, W.J., 1998. The logical problem of language acquisition in optimality theory. In: Barbosa, P., Fox, D., Hagstrom, P., McGinnis, M., Pesetsky, D. (Eds.), *Is the Best Good Enough? Optimality and Competition In Syntax*. MIT Press, Cambridge, MA, pp. 399–420.
- Revithiadou, A., Tzakosta, M., 2004. Alternative grammars in acquisition: markedness- vs. faithfulness-oriented learning. In: *Proceedings of the BUCLD28*, online supplement. <http://128.197.86.186/posters/revithiadou-BUCLD2003.pdf>.
- Roark, B., Demuth, K., 2000. Prosodic constraints and the learner's environment: a corpus study. In: Howell, S.C., Fish, S.A., Keith-Lucas, T. (Eds.), *Proceedings of the BUCLD 24*. Cascadilla Press, Somerville, MA, pp. 597–608.
- Rose, Y., 2000. *Headedness and prosodic licensing in the L1 acquisition of phonology*. Ph.D. Dissertation. McGill University.
- Smith, N.V., 1973. *The Acquisition of Phonology: A Case Study*. Cambridge University Press, New York.
- Smith, J.L., 2000. Positional faithfulness and learnability in Optimality Theory. In: Daly, R., Rehl, A. (Eds.), *Proceedings of the ESCOL99*. CLC Publications, Ithaca.
- Smith, J.L., 2001. Lexical category and phonological contrast'. In: Kirchner, R., Pater, J., Wikely, W. (Eds.), *PETL6: Proceedings of the Workshop on the Lexicon in Phonetics and Phonology*. University of Alberta, Edmonton, pp. 61–72.
- Smith, J.L., 2002. *Phonological augmentation in prominent positions*. Ph.D. Dissertation. University of Massachusetts Amherst.
- Smolensky, P., 1996. On the comprehension/production dilemma in child language. *Linguistic Inquiry* 27, 720–731.
- Smolensky, P., Legendre, G., 2006. *The Harmonic Mind*. MIT Press, Cambridge.
- Steriade, D., 1999. Alternatives to the syllabic interpretation of consonantal phonotactics. In: Fujimura, O., Joseph, B., Palek, B. (Eds.), *Proceedings of the 1998 Linguistics and Phonetics Conference*. The Karolinum Press, pp. 205–242.
- Stemberger, J., Bernhardt, B., Johnson, C.E., 2001. "Regressions" ("u"-shaped learning) in the acquisition of prosodic structure. ROA-471.
- Stites, J., Demuth, K., Cecilia, K., 2004. Markedness versus frequency effects in coda acquisition. In: Brugos, A., Micciulla, L., Christine, E.S. (Eds.), *Proceedings of the BUCLD 28*. Cascadilla Press, Somerville, MA, pp. 565–576.
- Tesar, B., 1998. An iterative strategy for language learning. *Lingua* 84, 131–155.
- Tesar, B., 2007. A comparison of lexicographic and linear numeric optimization using violation difference ratios. Ms. Rutgers University.
- Tesar, B., Smolensky, P., 1996. Learnability in optimality theory (short version). Technical Report JHU-CogSci-96-2. Cognitive Science Department, The Johns Hopkins University.
- Tesar, B., Smolensky, P., 1998. Learnability in optimality theory. *Linguistic Inquiry* 29 (2), 229–262.
- Tesar, B., Smolensky, P., 2000. *Learnability in Optimality Theory*. MIT Press, Cambridge, MA.
- Tessier, A.-M., 2006. Stages of phonological acquisition and error-selective learning. In: Baumer, D., Montero, D., Scanlon, M. (Eds.), *Proceedings of WCCFL25*. Cascadilla Press, Somerville, MA, pp. 408–416.
- Tessier, A.-M., 2007. *Biases and stages in phonological acquisition*. Ph.D. Dissertation. University of Massachusetts Amherst.
- Velleman, S., Vihman, M., 2003. The optimal initial state. Ms. University of Massachusetts Amherst and University of Wales at Bangor.
- Zoll, C., 1998. Positional asymmetries and licensing. Ms. MIT. Available at <http://roa.rutgers.edu/files/282-0998/roa-282-zoll-4.pdf>.
- Zuraw, K., 2000. *Patterned exceptions in phonology*. Ph.D. Dissertation. UCLA.