Linguistic Optimization*

Joe Pater, Rajesh Bhatt, and Christopher Potts

September 22, 2007

Abstract

Optimality Theory (OT) is a model of language that combines aspects of generative and connectionist linguistics. It is unique in the field in its use of a rank ordering on constraints, which is used to formalize optimization, the choice of the best of a set of potential linguistic forms. We show that phenomena argued to require ranking fall out equally from the form of optimization in OT's predecessor Harmonic Grammar (HG), which uses numerical weights to encode the relative strength of constraints. We further argue that the known problems for HG can be resolved by adopting assumptions about the nature of constraints that have precedents both in OT and elsewhere in computational and generative linguistics. This leads to a formal proof that if the range of each constraint is a bounded number of violations, HG generates a finite number of languages. This is nontrivial, since the set of possible weights for each constraint is nondenumerably infinite. We also briefly review some advantages of HG.

1 Introduction

Generative linguistics (GL) aims to formally characterize the set of possible human languages. For example, many languages place stress, or accent, on the first syllable of every word, and many stress the last syllable. No known human language measures the distance from each edge of the word and places an accent as close as possible to the middle. A successful generative account of stress provides a formal system that can express the first two patterns, but not the last one (see e.g., Hayes 1995; Gordon 2002). This is challenging because a system generating the last pattern is expressible with simple mathematics, and many real stress patterns may seem much more complex. Linguistic connectionism (LC) provides a theory of language based on mathematical analogues of neural activity (e.g., Rumelhart and McClelland 1986; Smolensky and Legendre 2006). The aims of linguistic connectionism are diverse, but there is a central focus on the modeling of cognition and learning. These two theories are often thought incompatible, since some versions of LC can learn and represent many implausible linguistic patterns (Pinker and Prince 1988; see Dresher 1994 for stress examples in a critique of Gupta and Touretzky 1994, and see relatedly Prince 1993; Goldsmith 1994). From the perspective of GL, such a theory would be insufficently *restrictive*.

By combining aspects of GL and LC, Optimality Theory (OT; Prince and Smolensky 1997, 2004) has provided a novel framework for linguistic analysis that has seen broad application, especially in phonology (see McCarthy 2002

*This paper has benefitted tremendously from discussion with a number of people. We would like to thank the participants at conferences and workshops where portions of it were presented, which took place at the University of Massachusetts, Amherst, the University of Orlé ans, the University of Chicago, Utrecht University, the Johns Hopkins University, the University of Tromso, the University of Amsterdam, the University of Nantes, and Stanford University. We would like to especially acknowledge the contributions of Adam Albright, Michael Becker, Paul Boersma, Hamida Demirdache, Kathryn Flack, Edward Flemming, Robert Frank, John Goldsmith, Karen Jesney, René Kager, Shigeto Kawahara, Shakuntala Mahanta, John McCarthy, Alan Prince, Curt Rice, Jason Riggle, Paul Smolensky, Anne-Michelle Tessier, Colin Wilson, and Kie Zuraw, though we hasten to add that this list is incomplete, and that remaining errors are ours alone. Portions of this research were supported by a Faculty Research Grant from the University of Massachusetts, Amherst, and by the Nederlandse Organisatsie voor Wetenschappelijk Onderzoek under NWO grant B 30-657.

for an overview). OT models linguistic systems in terms of constraints that place restrictions on the forms of the language. While there are other constraint-based theories in GL, OT is unique in having constraints that conflict, and a formal mechanism for resolving that conflict. In OT, the choice between linguistic representations is made by a ranking of constraints. When two constraints conflict in their preferences, the higher ranked constraint determines the outcome. This is an optimization system in that it chooses the best alternative from a space of possibilities.

In proposing constraint ranking, Prince and Smolensky depart from Harmonic Grammar (HG; Smolensky and Legendre 2006), the original fusion of LC and GL (see also Goldsmith 1990, 1993). In HG, optimality is defined using a harmony function in which the relative strength of constraints is expressed by coefficients, or weights. Prince and Smolensky (1997: 1608, 2004: 236) claim that using the harmony function would result in a theory that fails to distinguish properly between possible and impossible languages, but they do not present an example of an unattested linguistic pattern generated by weighting but not ranking.

In section 2, we discuss the linguistic patterns that Prince and Smolensky present as motivating the use of ranking. We show that they are also captured by an optimization system that uses weighting. This is partly due to the fact that an OT analysis of any linguistic pattern can be translated into an HG one (Prince and Smolensky 2004: 236; Prince 2002a; Legendre et al. 2006b). It is also due to the fact that many patterns that OT predicts are impossible are equally impossible in HG. Prince and Smolensky (1997: 1606) draw a broad generalization about linguistic patterns: that they do not display the effects of cumulative constraint violation. This might be thought to argue for ranking, since the preferences of a higher ranked constraint are independent of the degree of violation of lower ranked ones. In weighted optimization, a constraint with greater weight does not necessarily determine the outcome; so-called gang effects are possible. However, in an example of pairwise constraint conflict Prince and Smolensky provide, and in many others, avoiding a violation of one constraint entails a violation of the other. In this situation, weighting and ranking are equally insensitive to cumulativity.

Section 3 provides some evidence of the absence of cumulative effects of constraint violation by drawing on critiques of Smolensky's (2006) elaboration of OT, which introduces a mechanism of constraint conjunction. This theory predicts a number of unattested cumulative interactions. We show that unlike constraint conjunction, weighted optimization imposes inherent restrictions on *co-relevance* and *locality* in cumulative interactions (McCarthy 2003a), and does not generate the unattested cases.

Section 4 turns to a problem for HG raised by Legendre et al. (2006b). They provide an example of an interaction between a pair of OT constraints that produces implausible results in HG. Their example involves a type of constraint is controversial in OT (gradient Alignment; cf. McCarthy 2003b) and elsewhere in linguistics: it counts the distance of a stressed syllable from the edge of a word. We hypothesize that the source of the problem is not constraint weighting, but rather assumptions about the nature of constraints that are specific to OT. In this theory, constraints are global in their scope of evaluation, and there is therefore no upper limit on the number of violations that a constraint can assign. In many other approaches to GL and computational linguistics, including some versions of OT, constraints evaluate locally, and have an upper bound on the number of violations they assign.

Legendre et al. (2006b) also point out that in their example, the interaction of the two constraints in HG yields an infinite set of possible languages. We show that, if the set of violation marks is finite, HG yields a finite set of languages. The proof is described in section 5, and the proof itself is in an appendix.

Section 6 briefly discusses two arguments for the replacement of OT ranking with HG weighted constraints: (i) the resulting theory can be extended relatively straightforwardly to deal with various types of linguistic gradience, and (ii) it can make use of existing, well-understood algorithms for modeling learning and for other computational implementations.

2 Optimization with Weighted Constraints

This section introduces the notion of optimality in HG, and shows how optimization with weighted constraints yields some of the same effects as OT's ranked constraints. We present the central definitions in section 2.1, and then move to some illustrative applications in section 2.2. Section 2.3 discusses the difference between HG weighting and OT ranking: that only weighting can produce cumulative constraint interaction. Section 2.4 shows how optimization imposes inherent restrictions on cumulativity.

2.1 Optimality in HG

OT formalizes optimization as a choice amongst candidate mappings from one level of linguistic representation to another. In this section, we discuss the fundamentals of this formalization and how the harmony function can be used in such a system. Fuller definitions are given in the appendix.

In OT, the input and output are structures at two levels of linguistic representation. In the phonological examples we discuss in this paper, the input and the output are representations of the sound structure of an utterance. The input representation consists of hypothesized stored lexical representations of the individual words (or morphemes). In the observed output representation, each form may be altered in various ways given the properties of that utterance. For a given input structure I, the *candidate set* $\mathcal{A}_I$ for I is the set of all candidates with input I. This candidate set is produced by a function Gen, which associates the input with a set of potential outputs to produce a set of input–output pairs. Although OT uses ranked constraints to find the optimal candidate in a candidate set, Prince and Smolensky (2004: 236) point out that optimality could also be defined in terms of weighted constraints, as in HG. In HG, a candidate's harmony is the sum of its weighted violation counts, as defined in (1). Following Smolensky and Legendre (2006), we define this as a relation between a candidate, a set of n constraints C_n, and a weighting vector $W_n = w_1 \dots w_n$.

(1) $\mathcal{H}_{C_n,W_n}(A) = w_1(c_1(A)) + \dots + w_n(c_n(A))$

Here, each $c_i \in C_n$ is a function from candidates into integers. Following Legendre et al. (2006b), we convert OT violation marks to corresponding negative integers, thereby allowing for the possibility that some constraints might reward candidates by mapping them to positive numbers.

Optimality is always relative to candidate sets, and can be defined as the candidate with maximal *harmony*, as in (2) (Legendre et al. 2006b, and other papers in Smolensky and Legendre 2006). We will work with this definition of optimality throughout the paper, which we refer to as the *harmony function*.

(2) A candidate $A = \langle I, O \rangle$ is *optimal* for constraint set $C = \{c_1 \dots c_n\}$ and weighting $W_n = w_1 \dots w_n$ iff

$$\mathcal{H}_{C_n,W_n}(A) \geqslant \mathcal{H}_{C_n,W_n}(A')$$

for every $A' \in \mathcal{A}_I$.

We typically present HG evaluations in OT-style tableaux like the one in (3). Individual tableaux present (partial) candidate sets, so we specify the input they all share in the upper left, with the outputs listed below it. The top row provides the constraint weights, and the rightmost column provides the harmony score of each candidate.

(3) A weighted constraint tableau

	Weight	1.5	1	$\mathcal{H}$
	In Input_1	Constraint A	Constraint B	
☞	$\text{Output}_{1,1}$		-1	-1
	$\text{Output}_{1,2}$	-1		-1.5

Tableau (3) shows two candidates. They differ in whether they violate each constraint. The first candidate is optimal: the constraint it violates has the lower weight, so it has the higher negative score. The optimal mapping is indicated with the pointing finger.

We turn now to some examples of how differences in constraint weighting can yield linguistic differences.

2.2 Constraint weighting in linguistic analysis: an example

The next two tableaux present constraint weightings that would be appropriate for languages in which there are no voiced obstruents (e.g., [b], [d], [g], [z]) in syllable-final coda position (e.g., German, Dutch, Russian) and those in which such consonants are permitted (e.g., French, English, Spanish). The constraint *Coda-Voice penalizes voiced obstruents in coda position, and Ident-Voice penalizes mismatches in voicing between input and output (McCarthy and Prince 1999). If the weight of *Coda-Voice is greater than that of Ident-Voice, codas are produced as voiceless. We will assume an extremely simple Gen function: it creates input–output pairs in which input voiced obstruents like /b/, /d/, /g/, and /z/ are mapped to their voiceless counterparts [p], [t], [k] and [s]. For any input, the candidate set consists of every mapping that is created by changing some set of its voiced obstruents into the corresponding voiceless ones.[1]

In our first example, for input /bad/, there are four such mappings. The one that changes voicing only in coda position is optimal: it has the highest *harmony*.

(4) Coda devoicing

	Weight	1.5	1	$\mathcal{H}$
	In bad	*Coda-Voice	Ident-Voice	
	bad	-1		-1.5
☞	bat		-1	-1
	pad	-1	-1	-2.5
	pat		-2	-2

With the weightings reversed, codas can be produced as voiced, and the optimal output for /bad/ is [bad]:

(5) Voiced codas allowed

	Weight	1.5	1	$\mathcal{H}$
	In bad	Ident-Voice	*Coda-Voice	
☞	bad		-1	-1
	bat	-1		-1.5
	pad	-1	-1	-2.5
	pat	-2		-3

[1] Technically, this would be formalized as a change from [+voice] to [-voice] in a feature theory such as that of Chomsky and Halle 1968, or as deletion of [voice] in a theory such as that of Ito and Mester 1986. The restriction to obstruents could be stipulated in the operation, or derived from constraint interaction.

So long as the weight of *Coda-Voice is greater than that of Ident-Voice, coda devoicing occurs. If Ident-Voice has a greater weight than *Coda-Voice, voiced codas surface in outputs. In this case, weightings that respect these two strict inequalities have the same effect as the corresponding rankings (≫ is the symbol used to indicate dominance in an OT ranking):

(6) a. w(*Coda-Voice) > w(Ident-Voice) = *Coda-Voice ≫ Ident-Voice

b. w(Ident-Voice) > w(*Coda-Voice) = Ident-Voice ≫ *Coda-Voice

Under the restriction that constraint weights are positive real numbers (Prince 2002a), another equivalence between HG and OT obtains: no weighting of this set of constraints can make the /bad/ → [pad] mapping optimal. This mapping is *harmonically bounded* by both of the first two candidates fo /bad/. In any case in which one candidate has a proper superset of the violation marks of another, the superset candidate can be picked by no ranking (Prince and Smolensky 2004), and no weighting with positive reals (Prince 2002a). In this example, this fits with what we find cross-linguistically: many languages have coda devoicing, but none seem to have across-the-board syllable-initial devoicing.

In the above tableaux, the final candidate [pat] is harmonically bounded too, but can emerge when we consider a slightly expanded constraint set. Many languages do lack voiced obstruents altogether, which is accounted for with a constraint *Voice-Obs. The following table shows how this constraint and Ident-Voice evaluate the candidates for [bad].

(7) Violation scores on *Voice-Obsand Ident-Voice

In bad	*Voice-Obs	Ident-Voice
bad	-2	
bat	-1	-1
pad	-1	-1
pat		-2

If w(*Voice-Obs) > w(Ident-Voice), then /bad/ →[pat] is optimal. There is still no weighting that would make [pad] the optimal output for /bad/.

The restriction of weights to positive reals is generally necessary if HG is to function as an OT-like theory of language typology, since allowing positive and negative weights for a constraint would allow the same constraint to both penalize and *reward* violations. The version of HG we will consider in this paper therefore limits weights to positive values, as in Prince 2002a (see also the limitation to non-negative reals in Keller 2006).

2.3 Cumulativity in HG vs. strict domination in OT

To illustrate the main difference between HG and OT, we now consider a slightly more complicated example in the phonology of voicing. In Japanese, only a single sound from the set of voiced obstruents ([b], [d], [g], [z]) is usually permitted in a word (see Ito and Mester 1986, 2003 for analyses in GL). This restriction is termed *Lyman's Law* in honor of its discoverer. In loanwords, however, pairs of voiced obstruents are permitted (Nishimura 2003; Kawahara 2006; all data are from the latter source):

(8) Violations of Lyman's Law in loanwords

bagii 'buggy'	bagu 'bug'
bogii 'bogey'	dagu 'Doug'
bobu 'Bob'	giga 'giga'

Japanese also has a restriction against voicing in doubled (geminate) obstruents. But again, in loanwords, these occur:

(9) Voiced/voiceless obstruent geminate near-minimal pairs in Japanese loanwords

webbu ‘web’	wippu ‘whipped (cream)’
sunobbu ‘snob’	sutoppu ‘stop’
habburu ‘Hubble’	kappuru ‘couple’
kiddo ‘kid’	kitto ‘kit’
reddo ‘red’	autoretto ‘outlet’
heddo ‘head’	metto ‘helmet’

However, when a word contains both a voiced geminate and a voiced obstruent, the geminate is optionally devoiced:

(10) Optional devoicing of a geminate in Lyman’s Law environment

guddo ~ gutto ‘good’	doggu ~ dokku ‘dog’
beddo ~ betto ‘bed’	baggu ~ bakku ‘bag’
doreddo ~ doretto ‘dredlocks’	budda ~ butta ‘Buddha’
baddo ~ batto ‘bad’	doraggu ~ dorakku ‘drug’
deibiddo ~ deibitto ‘David’	biggu ~ bikku ‘big’

Kawahara (2006) shows that the devoicing in (10) results in a fully voiceless geminate. According to Nishimura (2003) and Kawahara (2006), such devoicing is judged unacceptable in both (8) and (9).

In HG, this devoicing pattern can be analyzed as an instance of cumulative interaction between two constraints. One is the constraint against voiced obstruent geminates (*Vce-Gem), and the other is a constraint that allows only one voiced obstruent in a word (a geminate counts as one), which we will call *2-Voice (Ito and Mester 1986, 2003). This analysis is inspired by Nishimura’s (2003) application of Smolensky’s (2006) Local Conjunction theory, which we discuss below. In (11), the weighting of Ident-Voice is greater than that of each of *2-Voice and *Voice-Obs, so a pair of voiced obstruents is permitted (/bobu/ → [bobu]), as is a voiced geminate (/webbu/ → [webbu]). The summed weighting of *2-Voice and *Vce-Gem is greater than that of Ident-Voice, so the geminate devoices in the presence of another voiced obstruent (/doggu/ → [dokku]). Again, we assume that Gen produces candidate sets consisting of all changes of voiced obstruents to voiceless. We leave out candidates that would never beat their competitors with this set of constraints, such as /bobu/→ [popu], which is harmonically bounded on this constraint set by the candidates we provide. The mapping /bobu/ → [popu] would be optimal if w(*Voice-Obs) > w(Ident-Voice), but this weighting would result in the absence of voiced obstruents throughout the language, counter to fact.

(11) Japanese loanword devoicing as cumulative constraint interaction

a.

	Weight	1.5	1	$\mathcal{H}$
	In bobu	Ident-Voice	*2-Voice	
☞	bobu		-1	-1
	bopu	-1		-1.5
	pobu	-1		-1.5

b.

	Weight	1.5	1	$\mathcal{H}$
	In webbu	Ident-Voice	*Vce-Gem	
☞	webbu		-1	-1
	weppu	-1		-1.5

c.

	Weight	1.5	1	1	$\mathcal{H}$
	In doggu	IDENT-VOICE	*VCE-GEM	*2-VOICE	
	doggu		-1	-1	-2
☞	dokku	-1			-1.5

No OT ranking of these constraints can generate this pattern (see Nishimura 2003 and Kawahara 2006 for expansions of the OT constraint set that deal with these data). With IDENT-VOICE ranked above each of *VCE-GEM and OCP-VOICE, /doggu/ would map to [doggu]. This is a result of what Prince and Smolensky (1997, 2004) call the strict domination property of ranked constraints: if a constraint c dominates a set of constraints C, this domination is strict in that no number of violations of C can overturn the preference of c.

Prince and Smolensky (1997: 1608, 2004: 236) and Legendre et al. (2006b) take strict domination to be an important property of a theory of linguistic optimization. They claim that the greater power of weighting is excessive, that it predicts unattested types of languages. The main point of the present paper is that HG is less powerful than it might at first seem, so that it potentially provides a sufficiently restrictive theory of language.

Before turning to the restrictiveness results, it is worth emphasizing the following: any linguistic pattern that can be analyzed with some set of ranked constraints can also be analyzed in terms of weightings of the same constraints (Prince and Smolensky 2004: 236; Prince 2002a). This bears on an oft-cited observation, which Prince and Smolensky (2004: 144) make of their analysis of Lardil final vowel truncation (Hale 1973):

> The relative harmonies of .yi.li.yil.<i> (183 A.i) and .yi.li.yi.li. (183 A.ii) pointedly illustrate the strictness of strict domination. Fully parsed .yi.li.yi.li. is less harmonic than truncated .yi.li.yil.<i> even though it violates only one constraint, while the truncated form violates three of the four lower ranked constraints . . .

The candidate representations Prince and Smolensky are discussing are shown in (12), along with the pattern of constraint violations. The meanings of the constraints are not important for present purposes. The point is simply that while the optimality of [yi.li.yil<i>] can indeed be used to illustrate constraint ranking, it can equally be used to illustrate constraint weighting:

(12) A weighting capturing the effect of strict domination in Lardil

	Weight	4	1	1	1	$\mathcal{H}$
	In yiliyili	FREE-V	ALIGN	PARSE	NOCODA	
☞	yi.li.yil.<i>		-1	-1	-1	-3
	yi.li.yi.li.	-1				-4

So long as the weight of FREE-V is greater than the summed weights of ALIGN, PARSE, and NOCODA, [yi.li.yil<i>] will emerge as optimal. Readers interested in constructing a full analysis of Lardil in HG terms can consult Potts et al. (2007) for software assistance and a computer readable version of Prince and Smolensky's (2004) Lardil tableaux.

2.4 Optimization restricts cumulative interaction in HG

In making the case for strict domination, Prince and Smolensky (1997: 1604) state the following:

> In a variety of clear cases where there is a strength asymmetry between two conflicting constraints, no amount of success on the weaker constraint can compensate for failure on the stronger one.

As an example, they discuss the interaction of NoCoda and Parse: NoCoda demands that syllables end in a vowel, rather than a consonant, and Parse demands that input segments be parsed into output syllable structure. Prince and Smolensky (1997: 1606) state that

> No matter how many consonant clusters appear in an input, and no matter how many consonants appear in any cluster, [the grammar with NoCoda ≫ Parse] ... will demand that they all be simplified by deletion (violating Parse as much as is required to eliminate the occasion for syllable codas), and [the grammar with Parse ≫ NoCoda] ... will demand that they all be syllabified (violating NoCoda as much as is necessary). No amount of failure on the violated constraints is rejected as excessive, as long as failure serves the cause of obtaining success on the dominating constraint.

In many, if not all, cases of NoCoda and Parse interaction this is just as true of any HG weighting as it is of any OT ranking (see relatedly Prince 2002a).[2] For example, HG is equally insensitive to the number of codas in the input: if the weighting value of NoCoda is greater than Max (McCarthy and Prince's (1999) replacement for Parse), all potential codas will be deleted, and if the weighting value of Max is greater than NoCoda, they will all surface faithfully. One might imagine that the additive constraint interaction of HG could produce a pattern in which one coda is tolerated but a second potential coda is deleted. That this is impossible can be demonstrated by determining the weighting conditions needed to produce each outcome.

In the following tables, the constraint violations and intended optima are provided for a hypothetical language in which one coda is permitted but any second coda in a word is deleted. The input /bat/ has one potential coda, which is realized faithfully in the optimal [bat]. The sub-optimal [ba] violates Max, since an input consonant is absent in the output. The input /bantat/ has two potential codas (syllable-final [n] and word-final [t]), and one is deleted in the optimal form. Here, Gen produces candidates with any set of deletions of input segments. We omit harmonically bounded candidates in which deletion does not result in improvement on NoCoda. We also omit /bantat/ → [batat], which is equivalent to /bantat/ → [banta] on these constraints.

(13) No one-coda maximum in HG

a.

In bat	NoCoda	Max
☞ bat	-1	
ba		-1

b.

In bantat	NoCoda	Max
bantat	-2	
☞ banta	-1	-1
bata		-2

The inequalities in (14) correspond to each of the above tables: for the intended optima to beat each of their competitors, these conditions must hold. In (14 b.), the inequality is the one required for the indicated mapping to beat /bantat/ → [banta]: the shared NoCoda violation in [banta] and [bantat] is irrelevant. These inequalities can be termed *weighting conditions* (cf. Prince's (2002b) OT ranking conditions). Because the weighting conditions are contradictory, no HG grammar can yield the indicated optima in (13).

[2] On some definitions of NoCoda it is possible to produce implausible patterns in HG. See the discussion surrounding (38) in section 4.5. With the definition of NoCoda Prince and Smolensky were working with, such cases might be taken as an argument for strict domination; even so, the argument is less general than they imply.

(14) a. /bat/ → [bat]: w(Max) > w(NoCoda)

b. /bantat/ → [banta]: w(NoCoda) > w(Max)

One can further note that deletion of only a single coda (i.e., [banta] from /bantat/) cannot be made to beat both of the competing mappings. With w(Max) > w(NoCoda), the mapping with no deletion is optimal (i.e., [bantat]); with w(NoCoda) > w(Max), deletion of both codas is (i.e., [bata]).[3]

This result would not hold if we set a numerical cut-off, or threshold, on well-formedness. For example, if we defined well-formed mappings as those with harmony above −1.5, we could produce the one-coda maximum, as shown in (15).

(15) Evaluation of mappings with cut-off of -1.5

	Weight	1	$\mathcal{H}$
		NoCoda	
√	/bat/ → [bat]	-1	-1
√	/bantat/ → [banta]	-1	-1
*	/bantat/ → [bantat]	-2	-2

An optimization system seeks the best outcome for each candidate set and does not impose this sort of threshold, and thus does not yield this sort of limit on the number of codas in a word.

This restrictiveness result is positive insofar as the one-coda maximum is unattested. In the above quotation, Prince and Smolensky appear to be implying that this is the case, and we know of no counter-examples. In the next section, we discuss the (lack of) evidence for cumulativity in the phonological literature. For now, it suffices to note that the absence of cumulativity in Prince and Smolensky's NoCoda/Parse example falls out from optimization, and cannot be used without elaboration as evidence for strict domination.

Prince (2002a) points out that the NoCoda/Parse interaction is an example of what he calls an "Anything Goes" system: to get an HG weighting to produce the same result as an OT ranking, we just have to replace the OT ranking relation "≫" with the numerical greater than relation ">", and assign as a weighting any set of positive real numbers that satisfies the demands of the resulting inequality. Prince's (2002a) concern is the translation of OT rankings to HG weightings. He does not comment on the implications for the typological predictions of HG or for Prince and Smolensky (1997: 1604) argument for strict domination.

3 Locality and co-relevance in HG cumulativity

Smolensky (2006) proposes a version of OT that includes a mechanism of local constraint conjunction (henceforth OT-LC), which allows OT to capture cumulative interactions between constraints. Smolensky defines local conjunction as follows (p. 43):

(16) Local conjunction within a domain D

*A $\&_D$*B is violated if and only if a violation of *A and a (distinct) violation of *B both occur within a single domain of type D.

Conjunction yields a new constraint that is violated iff its conjuncts are violated. When constraints are conjoined with themselves, the result is referred to a self-conjunction.

[3] Deletion of a single coda can be made optimal if elaborations of the theory discussed in sections 4.4 and 6 are adopted. If evaluation is local and probabilistic, then [bata], [banta], [batat] and [bantat] can be in free variation as outputs for /bantat/. See Riggle and Wilson 2005 for evidence that this sort of pattern is attested.

Critiques of local conjunction have pointed out that many of the cumulative constraint interactions that OT-LC produces do not exist in the phonologies of the world's languages (see esp. McCarthy 1999, 2003a). Here we demonstrate that HG is a restrictive theory of cumulative constraint interaction by providing two examples of unattested patterns that are generated by OT-LC but not HG.

Before presenting the problem cases, we will first show how two attested phonological patterns that we have already discussed can be analyzed in terms of cumulative constraint interaction in OT-LC. First, the pattern of final consonant devoicing can derived from the conjunction of NoCoda with *Voice-Obs in the domain of a single segment (Ito and Mester 2003). Here we adopt OT tableaux conventions: violations are indicated with an asterisk, and a fatal violation is highlighted with an exclamation mark. A solid line separates constraints that are crucially ranked, while a dotted line separates constraints whose ranking is indeterminate. A candidate Can_1 is eliminated iff there is another candidate Can_2 that has fewer violations of a constraint c and there is no constraint ranked above c on which Can_2 has fewer violations than Can_1. The exclamation mark indicates the constraint violation that causes elimination. The optimum is any candidate that is not eliminated.

(17) Local conjunction analysis of final devoicing

In bad	NoCoda &$_{Seg}$*Voice-Obs	Ident-Voice	NoCoda	*Voice-Obs
☞ bat		*	*	*
bad	*!		*	**
pat		**!	*	
pad	*!	*	*	*

The voiced obstruent in initial position remains voiced because Ident-Voice ranks above *Voice-Obs, the constraint that penalizes these segments. A voiced obstruent in coda position, however, violates both NoCoda and *Voice-Obs, and is therefore targeted by the conjoined NoCoda &$_{Seg}$*Voice-Obs. Because this constraint ranks above Ident-Voice, voicing is lost in coda position. Standard OT and HG cannot reduce coda devoicing to the effect of NoCoda and *Voice-Obs, and must postulate an independent constraint like *Coda-Voice above (but see Lombardi 1999 for an alternative analysis, and see Steriade 1999 for arguments that this sort of positional restriction is not due to the cumulative effect of a syllabic constraint like NoCoda and a segmental constraint like *Voice-Obs).

The other phonological pattern we have already discussed that can be dealt with in this way is Lyman's Law in Japanese (Alderete 1997; Suzuki 1998; Ito and Mester 2003). The self-conjunction of *Voice-Obs over the domain of the word penalizes any word that contains two voiced obstruents:

(18) OT-LC analysis of Lyman's Law

In bada	*Voice-Obs &$_{Word}$*Voice-Obs	Ident-Voice	*Voice-Obs
☞ bata		*	*
bada	*!		**
pata		**!	

The candidate with two output voiced obstruents violates *Voice-Obs &$_{Word}$*Voice-Obs. Because it is ranked above Ident-Voice, devoicing of one of the obstruents is preferred over the faithful realization. Because Ident-Voice is ranked above *Voice-Obs, devoicing of the second obstruent is ruled out. Again, standard OT and HG must postulate an independent constraint, like our *2-Voice above, to capture this pattern.

One of the problematic predictions of OT-LC can be illustrated by combining these two analyses. In the following tableau, we make use of a conjunction of NoCoda and *Voice-Obs over the domain of the word. As (19) shows, when this constraint is ranked above Ident-Voice, words that contain both a voiced obstruent and a coda are ruled out. With Ident-Voice ranked above *Voice-Obs, a voiced obstruent in a word with no coda (e.g., [ba]) will be allowed.

(19) OT-LC analysis of unattested cumulative pattern

In bat	NoCoda &$_{\text{Word}}$*Voice-Obs	Ident-Voice	NoCoda	*Voice-Obs
☞ pat		*	*	
bat	*!		*	*

This long-distance devoicing in the presence of a coda is clearly unattested: bans against codas are common, as are bans against voiced obstruents, but as far as we know, nothing like this pattern has ever been reported in the phonological literature. A range of similarly implausible cases can be produced by conjoining different constraints over the domain of the word. This seems to indicate that restrictions against multiple instances of a segment type, like Lyman's Law, are best understood as being due to constraints like *2-Voice that directly penalize these configurations, rather than as responses to a general constraint schema that can potentially penalize any instance of multiple constraint violation (see also Ito and Mester 2003: 59-61 for related discussion).

To show that this pattern is not produced in HG, we can consider the weighting conditions that would be required to make /bat/ → [pat] and /ba/ → [ba] jointly optimal. The pattern of constraint violations represented with integers is shown in (20).

(20) Constraint violations for unattested pattern

In bat	NoCoda	*Voice-Obs	Ident-Voice
bat	-1	-1	
☞pat	-1		-1

In ba	NoCoda	*Voice-Obs	Ident-Voice
☞ba		-1	
pa			-1

To make /bat/ → [pat] optimal, the weight of *Voice-Obs would have to be greater than that of Ident-Voice (the shared violation of NoCoda is irrelevant). For /ba/ → [ba], we would need a weighting that satisfies the opposite demand: w(Ident-Voice) > w(*Voice-Obs). This pattern can therefore be produced by no weighting of the constraints.

Even constraints conjoined over their smallest shared domain can produce implausible results. The following example involves a conjunction of *Voice-Obs with Agree-Place, another constraint whose effects are commonly observed cross-linguistically. Many languages require that adjacent consonants have the same place of articulation, or position of oral closure. The labial consonants [p], [b], and [m] have closure of the lips, while [t], [d], and [n] are articulated with the tip of the tongue. Agree-Place penalizes pairs of adjacent consonants like [bn] and [tm] that have different places of articulation. By conjoining these constraints over the domain of the segment that is targeted by both constraints, and ranking this constraint over Ident-Voice, a segment will lose voicing when it is part of a cluster that disagrees in place of articulation. If Ident-Voice is ranked above *Voice-Obs, voicing will be permitted in other environments.

(21) Local conjunction analysis of an unattested pattern

a.

In pob+ni	Agree-Place &$_{\text{Seg}}$*Voice-Obs	Ident-Voice	Agree-Place	*Voice-Obs
pob+ni	*!		*	*
☞ pop+ni		*	*	

b.

In pob+mi	Agree-Place &$_{\text{Seg}}$*Voice-Obs	Ident-Voice	Agree-Place	*Voice-Obs
☞ pob+mi				*
pop+mi		*!		

As far as we know, no language has this distribution of voiced obstruents. Again, this pattern cannot be produced by HG. For /pob+ni/ → [pop+ni], we need w(*Voice-Obs) > w(Ident-Voice) (the shared Agree-Place violation is irrelevant). For /pob+mi/ → [pop+mi], the contradictory inequality w(Ident-Voice) > w(*Voice-Obs) must hold.

In this case, the problem for the OT-LC theory of cumulative interaction is one of *co-relevance* (McCarthy 2003a). The voicing of a segment and its relationship with the following segment in terms of place of articulation are independent factors. In OT-LC, they are allowed to affect one another.

Some proposals aim to restrict the power of OT-LC by imposing restrictions on the types of constraints that can be conjoined (to solve co-relevance problems), or the domains over which they can be conjoined (to solve locality problems); Lubowicz (2005) provides a recent review. It is not our purpose here to judge the success of OT-LC in general or of these restrictions in particular. Instead, we seek only to point out that, unlike OT-LC, HG imposes inherent restrictions on co-relevance and locality. The source of these restrictions is that for cumulativity to have an effect in HG, the multiple violations in the sub-optimal candidate must be able to be traded off against a smaller number of violations in the optimum. In the Japanese loanword example, the violations of *2-Voice and *Vce-Gem could be avoided with a single violation of Ident-Voice. Unrelated constraint violations, and constraint violations in different parts of the string, cannot typically be resolved in a single way. While the generality of this statement is dependent both on the contents of the constraint set and on how the candidates are generated, for the cases we have discussed here, HG provides a more restrictive theory of cumulative constraint interaction than does OT-LC.

4 Locality and boundedness of violation

In section 4.1, we review the problem for HG raised by Legendre et al. (2006b). In section 4.2, we point out their example relies on a controversial type of constraint. This constraint is global in the sense that it assesses the relationship between elements of a representation at an arbitrary distance from one another. Such constraints are generally prohibited in GL, and are ruled out in some versions of OT. In section 4.3, we show that HG does still create implausible linguistic patterns with local constraints. We argue that these further problems are due to the way that constraints evaluate representations in standard OT, which is also global in that all aspects of the representation are assessed once and only once. This globality of evaluation creates problems in OT as well as HG. In sections 4.4 and 4.5, we show how adopting and extending a proposed solution to the OT problems can resolve these issues in HG. A result of imposing locality restrictions on HG and OT is that it becomes plausible that there is an upper bound on the degree of violation of HG/OT constraints. This boundedness assumption is key to the proof in section 5.

4.1 A locality problem for HG

Legendre et al. (2006b) provide an example of the interaction of two constraints that yields implausible results in HG, but not in OT. MainStressRight assigns a violation mark for each syllable between the main stress and the right edge of the word. StressHeavy demands that particular type of syllable, termed heavy, be stressed. In our example, heavy syllables are those that end in a coda, and light ones end in a vowel. The output candidates have stress on any one of the syllables, which is indicated with a preceding single quotation mark. These two constraints conflict whenever the rightmost heavy syllable is non-final: MainStressRight prefers final stress, and StressHeavy prefers stress on the heavy syllable. In (22), we see that if the weight of StressHeavy is greater than that of MainStressRight, stress will fall on a heavy syllable ([ban]) that is one syllable away from the edge, rather on a final light syllable ([la]).

(22) Stress on the heavy syllable

	Weight	1.5	1	$\mathcal{H}$
	In ban.la	StressHeavy	MainStressRight	
☞	'ban.la		-1	-1
	ban.'la	-1		-2.5

However, with the weighting in (22), if the heavy syllable is two syllables away from the right edge, final stress is preferred, as illustrated in (23).

(23) Stress on the final syllable

	Weight	1.5	1	$\mathcal{H}$
	In ban.ta.la	StressHeavy	MainStressRight	
☞	ban.ta.'la	-1		-1.5
	ban.'ta.la	-1	-1	-2.5
	'ban.ta.la		-2	-2

This sort of pattern is cross-linguistically attested, but as Legendre et al. point out, other weightings produce a range of unattested possibilities. Languages can be created that allow stress up to a maximum of *n* syllables from the word edge, where *n* is any integer. Languages exist that allow stress to fall only 0, 1, or 2 syllables from the right edge, but ones that place higher bounds on permissible distance from the edge do not seem to be attested. As Legendre et al. note, because there is no upper bound on the length of words, this pairwise constraint interaction produces infinitely many languages.

We hypothesize that the source of this problem is an assumption about the nature of constraints common to HG and OT, and not HG's use of weighted constraints, which is the cause Legendre et al. identify.

4.2 Locality in constraint formulation

The MainStressRight constraint in the above example counts the distance of a stressed syllable from the edge of the word. In constraint-based theories of stress outside of OT, constraints are local: they refer only to adjacent elements in a representation (e.g., McCarthy and Prince 1986; Halle and Vergnaud 1987; Hayes 1995). MainStressRight is one of a family of gradient Alignment constraints in the theory of McCarthy and Prince (1993). Gradient Alignment is problematic not only because it is formally atypical of linguistic constraints (Eisner 1998; Bíró 2003), but also because it produces implausible patterns in OT as well as HG (e.g., Kager 2001; Gordon 2002; McCarthy 2003b). In McCarthy's (2003b) revised theory of OT constraints, gradient Alignment is banned along with other non-local constraints. McCarthy proposes that all constraints must conform to the following schema:

(24) $*\lambda/\mathbb{C}$

For any λ satisfying condition $\mathbb{C}$, assign a violation mark

The element λ is the *locus of violation* of the constraint, and $\mathbb{C}$ is the *context*. All of the constraints we have discussed to this point (except gradient Alignment) can be formulated in these terms.[4] For example, the constraint Agree-Place from section 3 could be stated as in (25):

[4] For *2-Voice to be formulated to apply to adjacent constituents, there would have to be a level at two which voiced obstruents are adjacent. The standard account is to place the voicing feature on a separate representational plane (Ito and Mester 1986); see Ito and Mester (2003) for challenges for this account, and Suzuki 1998 for an alternative approach.

(25) AGREE-PLACE

For any consonant (λ) followed by a consonant with a different place of articulation ($\mathbb{C}$), assign a violation mark.

The locus of violation is a single linguistic element, or constituent. Here, it is a phonological segment. McCarthy 2003b: 80 cites Paul Smolensky (p.c.) as suggesting that the context could also be limited to a single constituent.[5] We adopt this restriction, which places strong locality limits in the statement of constraints.

By adopting this restriction, we eliminate Legendre et al.'s (2006b) example of an unattested HG system (see McCarthy 2003b for reanalysis of stress patterns without gradient Alignment). However, similar examples can be created with constraints that abide by this locality restriction. In the following section, we provide two examples, and show how these are the result of global evaluation, which also causes problems in OT.

4.3 Non-locality with local constraints in HG and OT

In this next example of an unattested pattern produced by HG, we use two local categorical constraints. The first is AGREE-ATR, a constraint that demands that two vowels have the same value for ATR (Bakovic 2000). ATR (Advanced Tongue Root) is a feature that characterizes the difference between vowels like English +ATR [i] (as in 'beet') and -ATR [I] (as in 'bit'; we use capital I instead of the standard phonetic transcription [ɪ] for legibility). In a number of languages surveyed by Bakovic and others, sequences of vowels must have the same value of this feature, undergoing a process usually termed ATR harmony. We formulate the constraint in the McCarthy (2003b) schema as follows:

(26) AGREE-ATR

For any vowel (λ) followed by a vowel with a different value of ATR ($\mathbb{C}$), assign a violation mark.

This constraint ignores intervening consonants, perhaps because vowels are represented on a separate plane. The second constraint, IDENT-ATR, penalizes each change in ATR value between input and output. In some ATR harmony systems [-ATR] vowels assimilate to [+ATR]. To simplify, our candidate set will display only changes from [-ATR] to [+ATR] (Bakovic (2000) uses an IDENT constraint that protects the [+ATR] vowels). The set of candidate mappings will consist of a given input paired with outputs that display all possible sets of changes from [-ATR] to [+ATR], from no change at all through to the change of all vowels.

In (27), we provide a schematic example of a how a weighting of AGREE-ATR above IDENT-ATR yields the attested pattern. The vowel we transcribe as [E] is a -ATR vowel (as in 'bet'), and [e] and [i] are [+ATR]. The plus sign shows the separation of the phonological string into morphemes. With w(AGREE-ATR) > w(IDENT-ATR), vowel harmony occurs. The tableaux in (27) show that forms with harmony are chosen over those that retain input ATR specifications.

(27) An attested vowel harmony system

Weight	1.5	1	$\mathcal{H}$
In mE+ti	AGREE-ATR	IDENT-ATR	
mE+ti	-1		-1.5
☞ me+ti		-1	-1

In the following tableaux we provide longer strings of syllables, and a weighting that produces an implausible system. In the first tableau, two [-ATR] vowels assimilate to the [+ATR] one. This is an attested pattern in that vowel harmony

[5] A slightly weaker version, and one with much precedent in phonological theory, would be to limit the context to stating the elements immediately preceding and following the locus of violation. The cumulativity of constraint violation in HG may allow the stronger version to be upheld, since a two-sided condition could be reduced to independent one-sided conditions (see Flemming 2001 for relevant discussion).

usually extends across the whole domain in which it applies (unless it is *blocked*; see below). Domains for vowel harmony are usually the root and some set of affixes, which make up a word. In our hypothetical example, there is a root /nEmE/, and an affix /ti/.

(28) Long-distance assimilation.

Weight	2.5	1	$\mathcal{H}$
In nEmE+ti	AGREE-ATR	IDENT-ATR	
nEmE+ti	-1		-2.5
nEme+ti	-1	-1	-3.5
nemE+ti	-2	-1	-6
☞ neme+ti		-2	-2

The following tableau shows the implausible pattern that this weighting produces. Here we have added one more [-ATR] vowel to the beginning of the input string. Now the candidate without change is optimal.

(29) An unattested blocking pattern

Weight	2.5	1	$\mathcal{H}$
In bEnEmE+ti	AGREE-ATR	IDENT-ATR	
☞ bEnEmE+ti	-1		-2.5
bEnemE+ti	-3	-1	-8.5
benEmE+ti	-2	-1	-6
bEneme+ti	-1	-2	-4.5
benEme+ti	-2	-2	-7
benemE+ti	-2	-2	-7
beneme+ti		-3	-3

In this example, the decision of whether harmony applies is being made based on the presence or absence of a vowel three syllables away from the "trigger" of harmony, the [+ATR] vowel [i]. As in the example provided by Legendre et al. (2006b), there are infinitely many possible variants of this pattern. Constraint weightings can yield vowel harmony patterns in which a string of n vowels will assimilate to an adjacent vowel, and in which a string of $n + 1$ vowels will fail to assimilate, where n is any integer.[6]

A clear cross-linguistic generalization about ATR harmony, and about harmony of other vowel features is that the sheer number of vowels preceding the trigger of harmony is irrelevant. As far as we know, no language has even the pattern where two vowels assimilates to a following one, and a string of three vowels does not, let alone any of the further patterns this system can produce. These patterns are highly non-local in that the decision of whether or not assimilation applies between a trigger and an adjacent potential target is dependent on the absence or presence of a vowel at an arbitrary distance.

This example shows that locality in constraint formulation is not a guarantee of local behavior in an HG system. The source of non-locality in this case is the globality of evaluation. In standard OT, a candidate set consists of all possible mappings from a given input to some output. The constraint set evaluates the candidate set and finds the optimal mapping. Under this view, all aspects of the representation are evaluated once and only once by the entire constraint set. In our HG example, the result is that parts of the representation at an arbitrary distance from one another can interact in a fashion that seems uncharacteristic of human language. This is also true of OT, as McCarthy (2003b;

[6]This infinite typology can also be produced if we allow the change from +ATR to -ATR as an operation in Gen alongside the -ATR to +ATR one. So long as there is a constraint that blocks this change, the weighting of this constraint can always be made high enough to rule it out.

2006; 2007) and Wilson (2003) have shown. To illustrate a globality problem in OT noted by McCarthy (2003b), we will further elaborate our hypothetical ATR harmony example.

In many ATR harmony systems, certain vowels resist change, with the result that some vowel sequences that disagree in ATR specification occur in the surface forms of the language, even though the language generally requires agreement. Usually, this is because a -ATR vowel has no +ATR counterpart. In the vowel system of our hypothetical language, we will assume that the -ATR vowel [A] has no +ATR counterpart [a]. In transcriptions of real languages, [a] is usually -ATR, but we adopt a uniform system of indicating -ATR vowels with capital letters. We introduce a constraint **a* that assigns a violation mark to an instance of the +ATR vowel [a].

The OT ranking producing harmony with blocking would be **a* ≫ AGREE-ATR ≫ IDENT-ATR. We again adopt OT tableaux conventions. In the tableaux in (30), we see how the ranking AGREE-ATR ≫ IDENT-ATR yields harmony, and **a* ≫ AGREE-ATR yields blocking.

(30) Assimilation and blocking in OT

a.

In nE+ti	**a*	AGREE-ATR	IDENT-ATR
nE+ti		*!	
☞ ne+ti			*

b.

In nA+ti	**a*	AGREE-ATR	IDENT-ATR
☞ nA+ti		*	
na+ti	*!		*

So far, the result is parallel to what is found cross-linguistically. However, for an input in which a vowel that can undergo harmony is preceded by the blocking vowel, we get an unattested result. As shown in (31), the candidate without assimilation is optimal. McCarthy (2003b) labels this the 'sour grapes' problem: if AGREE-ATR cannot be completely satisfied, it has no effect at all. This is again a highly non-local effect: it will occur when the blocking vowel [a] is separated by any number of potential undergoers like [E] from the trigger [i].

(31) The sour grapes problem for global OT

In nAmE+ti	**a*	AGREE-ATR	IDENT-ATR
☞ nAmE+ti		*	
nAme+ti		*	*!
namE+ti	*!		*
name+ti	*!		**

Along with ATR harmony, assimilation of other vowel features is extremely common cross-linguistically, and in many of these vowel harmony systems there is a subset of vowels and/or morphemes that fails to assimilate. In all of these cases, assimilation proceeds up to the blocking vowel, as in the failed candidate [nAme+ti] in tableau (31). The problem is that the desired result is harmonically bounded by the sour grapes candidate, [nAmE+ti] in (31). To address this problem, McCarthy (2003b, 2004) proposes replacements for AGREE constraints, while Wilson (2003, 2006b) proposes revisions to the candidate generation and evaluation process. Our own proposed solution is inspired by Wilson's (2003; 2006b) approach, but is formally an extension of McCarthy's (2006) attack on other locality problems in OT. Before turning to our proposal, we will finish this section by showing how one of McCarthy's (2006) examples raises a similar problem for HG. This also serves to demonstrate that globality of evaluation is a general problem for both HG and OT.

This time, the unattested pattern is long-distance metathesis. Metathesis is a change in the ordering of segments. Languages sometimes employ local metathesis to improve syllable structure. For example, the change in the order of

[t] and [a] in (32) avoids a violation of NoCoda. The conflicting constraint is Linearity (McCarthy and Prince 1999), which assigns a violation mark to each pair of segments whose precedence relation is reversed between input and output.

(32) An attested pattern of metathesis

	In at+ki	NoCoda	Linearity
☞	taki		*
	atki	* !	

Under standard OT's global evaluation this same ranking yields the unattested long-distance metathesis, as shown in (33).

(33) An unattested pattern of metathesis

	In unomelik+at	NoCoda	Linearity
☞	tunomelika		*********
	unomelikat	* !	

Moving the final [t] of the input to initial position in the output avoids a violation of NoCoda, and incurs 9 violations of Linearity because the precedence relation of [t] and each of the other 9 segments has been changed. According to McCarthy (2006; 2007), this pattern of long-distance metathesis is unattested. In HG the problem is amplified in that we get many more unattested patterns, and again an infinite set of possibilities. With these constraints in HG, languages can be created in which a coda will be avoided by incurring n violations of Linearity, but not $n + 1$, where n is any integer.

Global evaluation thus leads to implausible systems in both HG and OT. Arguments for bounded domains of evaluation (i.e. against global evaluation) also come from theoretical perspectives other than HG and OT and from empirical domains other than phonology. A central insight in syntactic theory is that syntactic processes are local i.e. they operate over domains of bounded size; the set of elements that any given element can be related to is always a small finite set. On the other hand, at least on the surface, we know that any domain can be made arbitrarily large (e.g. *the book*, *the book [that won the Pulitzer]*, ...). This unboundedness, which is introduced by recursion, is real but it does not actually influence the locality properties addressed above. The unboundedness introduced by recursion can be factored out leaving us with elementary objects of bounded size - the 'kernel sentences' of early generative grammar (Chomsky1957; 1985), the elementary trees of Tree Adjoining Grammars (Joshi et al. 1975), or the phases of Minimalism (Chomsky2000; 2001). These different approaches have in common the property that designated parts of a larger syntactic derivation are bounded in size and have to satisfy syntactic wellformedness and optimation/economy conditions individually. Such conceptions of grammar eliminate the need for global evaluation. Much like the case with HG and OT, the possibility of global evaluation, unless otherwise constrained, will predict implausible patterns of unattested non-local interactions.

Thus, the alternative to global evaluation is to have constraints evaluate smaller portions of the representation, that is, to have local evaluation. In section 4.4, we discuss an alternative approach to OT candidate generation and evaluation termed Harmonic Serialism by Prince and Smolensky (2004: 95). As McCarthy (2006; 2007) shows, Harmonic Serialism entails some types of local evaluation: we discuss his solution to the OT long-distance metathesis problem, which also rules out the related unattested HG systems. In section 4.5, we propose an amendment to Harmonic Serialism that imposes further locality restrictions, which resolves some outstanding problems for this theory, permits a resolution of the OT sour grapes problem, and rules out the unattested vowel harmony systems in HG.

We emphasize, however, that this is far from the only method of formalizing local evaluation in OT and HG. Tesar (1995b), Eisner (1997), Wilson (2006b), Bíró (2006) and others provide computational implementations of evaluation in OT that are local in various ways (see Riggle 2004 for a very useful review of much of this literature). The connectionist implementations of HG in Legendre et al. (2006b) and Soderstrom et al. (2006) are also non-global in that the representation is gradually altered. In addition, most, if not all, research in GL outside of OT assumes some form of local evaluation in which changes to a representation (rules or transformations) apply iteratively in local domains. The combination of iterative rules with constraints is often referred to in phonology as a constraints-and-repairs theory (Paradis 1988). Goldsmith (1993) proposes a formalization of the interaction of rules and constraints which is closely related to Harmonic Serialism. McCarthy (2002) and Prince and Smolensky (2004) provide extensive references to work in other constraint-based theories in GL, especially in phonology, and comparison with OT; see also McCawley (1968); Chomsky and Lasnik (1977); Johnson and Postal (1980); Kaplan and Bresnan (1982); Pullum and Scholz (2001) in syntax.

Local evaluation means that the constraints' ranges can be finite, in fact extremely small, as in many frameworks outside OT. In global OT, there can be no upper bound on the number of violations of a constraint, since the entire representation is evaluated in one fell swoop, and there is no upper bound on the length of a representation. In GL outside of OT, there are typically only two states for a constraint: violated or satisfied. This is true also of some finite-state approaches to constraint-based linguistics (see e.g., Frank and Satta 1998 and Karttunen 1998 in OT). Boundedness may be maintained even if constraints make more than a binary distinction, for instance if they assign negative scores for violation and positive ones for satisfaction (Legendre et al. 2006b), or if they make a finite number of scalar distinctions (e.g., H-NUC in Prince and Smolensky 2004).

4.4 Harmonic Serialism and globality problems in OT

Prince and Smolensky (2004: 95) briefly discuss an alternative to standard OT's mode of candidate generation and evaluation, which they term Harmonic Serialism (HS). In HS, only a single change to a representation is made at a time. This change (or a set of candidates with single changes) is then submitted to the constraint system to determine if it improves harmony relative to the unchanged representation. This process iterates until change ceases to improve harmony, at which point the *derivation* terminates.

We will now present McCarthy's (2006; 2007) HS solution to the long-distance metathesis problem. We assume that the metathesis operation permutes the order of any single pair of segments, so the first step of the derivation for the input string in (33) would produce candidates like those shown in as in (34). These candidates illustrate the fact that it is impossible to resolve the NOCODA violation with a single metathesis operation. In particular, the final candidate, which switches the positions of the final consonant and vowel, results in a new violation of NOCODA, so long as the [k] appears in coda position. As such, this first step in the derivation would also be the last. Since harmony has not improved, the derivation terminates without having produced metathesis.

(34) The first step in the HS derivation

In unomelik+at	NOCODA	LINEARITY
☞ unomelikat	*	
nuomelikat	*	*!
uonmelikat	*	*!
⋮		
unomelikta	*	*!

McCarthy's solution to the long-distance metathesis problem holds equally for HG as it does for OT, since harmonic improvement can be defined either in OT or HG terms. The infinite typology discussed beneath (33) is eliminated under HS because only a single LINEARITY violation is considered at a time. To produce the infinite typology, the number of times that LINEARITY is violated in long-distance movement must be counted.

It may in fact be true of all Faithfulness constraints in HS that in each evaluation, each constraint is violated maximally once. Faithfulness constraints are those constraints that specify the relationship between input and output: along with LINEARITY, the ones we have used include the IDENT constraints and MAX. Whether every operation leads to at most one violation of each of these constraints will depend on the formulation of the operations and the constraints, but this is true of all of the examples of operations and Faithfulness constraints in this paper. McCarthy (2007) in fact takes it as definitional of a step in a HS derivation that it violates just one 'basic' Faithfulness constraint.

HS thus imposes a locality restriction on OT and HG: for a series of operations to achieve a non-local goal, they must meet intermediate local goals along the way. In the next section, we propose a further locality restriction on HS, which limits evaluation to the structures manipulated by an operation. This solves some outstanding problems for HS, and allows a successful account of the vowel harmony patterns discussed above. It has the further result that constraints other than Faithfulness should have an upper bound on the number of violations.

4.5 Local harmonic serialism

One problem for HS relates to the NOCODA constraint, which either assigns a violation mark for every syllable that ends in a consonant (Prince and Smolensky 2004), or for every consonant that is final in some syllable (McCarthy 2003a). In standard OT and HG, and in HS, these are equivalent. The problem for HS arises when we consider syllables that end in one consonant and in two or more. With NOCODA ranked above MAX, a single consonant is deleted, but a pair of consonants is retained. This unattested pattern arises because of the limitation to a single operation, in this case, to a single deletion of a consonant. We leave out candidates in which a consonant is deleted from onset position, which will never result in harmonic improvement. The mapping /bant/ → [bant] is optimal because the entire coda cannot be deleted in a single step (that is, [ba] is not an output candidate in the second tableau).

(35) An unattested pattern of coda deletion in OT with HS

In ban	NOCODA	MAX
ban	* !	
☞ ba		*

In bant	NOCODA	MAX
☞ bant	*	
bat	*	*!

There are a number of conceivable solutions to this problem. For example, NOCODA could be reformulated to assign a violation mark to every segment in coda position, so that each deletion leads to harmonic improvement. Rather than changing the definition of NOCODA, we propose a general revision to the way that constraints assign violations in HS.

The constraints that evaluate output structures in OT are usually termed Markedness constraints. Examples of such constraints in this paper include NOCODA, the constraints on voicing like *VOICE-OBS, and the AGREE constraints. We propose that a Markedness constraint assigns a violation *iff* its locus of violation is altered by the operation that has initiated the current evaluation. Locus of violation is a designated element of a representation in McCarthy's (2003b) theory of OT constraints, discussed in section 4.2. We term the version of HS that assigns violations in this fashion Local Harmonic Serialism (LHS).

McCarthy's (2003a) definition of NoCoda designates a consonant as the locus, and the context as syllable-final position. In LHS, a violation of NoCoda is assigned *iff* the consonant affected by an operation is the locus of violation.[7] Thus, when a candidate set is formed by deleting the [n] of /bant/, only the [n] violates NoCoda, as in (36).

(36) The first step in the LHS derivation

	In bant	NoCoda	Max
	bant	* !	
☞	bat		*

We assume that the candidate set in LHS is formed by applying an operation to each eligible element in turn, and evaluating whether that change results in harmonic improvement. If the string is scanned from left-to-right, then the evaluation in (36) would have been the second one: deletion of [b] would not have resulted in Harmonic improvement, and so that change would be rejected. When Harmonic improvement is obtained, the new representation is taken as the input for further operations and evaluation. If a pass through the representation results in no change, the derivation terminates. This definition of how operations apply is somewhat underformalized, and does not specify how multiple operations interact, but will suffice for our goals and for the simple cases we consider here, which involve just a single operation.

The next step in the derivation we have been constructing for [bant] will involve the deletion of the [t] and evaluation of the result (again, deletion of [b] would not increase harmony).

(37) The second step in the derivation

	bant	NoCoda	Max
	bat	* !	
☞	ba		*

After this step, the derivation would terminate, since [b]-deletion would be rejected as non-harmonically improving.

Under this revised version of HS, we thus get the correct outcome: a pair of consonants will be deleted under the ranking NoCoda ≫ Max. It is worth noting that a revised approach to coda deletion is necessary not only for HS, but also for HG. If we adopted the standard OT interaction between NoCoda and Max in HG, we also predict the unattested pattern of deletion of only a single coda, as shown in (38).

(38) An unattested pattern of coda deletion in global HG

	Weight	1.5	1	$\mathcal{H}$
	In ban	NoCoda	Max	
	ban	-1		-1.5
☞	ba		-1	-1

	Weight	1.5	1	$\mathcal{H}$
	In bant	NoCoda	Max	
☞	bant	-1		-1.5
	bat	-1	-1	-2.5
	ba		-2	-2

[7] The tableau in (34), which demonstrates how long-distance metathesis is ruled out in HS, shows NoCoda violations assigned in the standard fashion. However, in LHS, the same basic point holds: metathesis can only occur if it increases harmony in each step. Even though the final candidate would not violate NoCoda, it would violate other constraints.

In the LHS analysis, this pattern would not arise, since only a single MAX violation is incurred at a time. This pattern would also be impossible in HG under the alternative solution to the HS NOCODA problem in which every segment in coda position violates the constraint.

The second, related problem for standard HS arises with constraints like AGREE-ATR. As shown in (39), when there is a string of -ATR vowels followed by a +ATR vowel, a single change from -ATR to +ATR does not lead to harmonic improvement. The mapping /nEmE+ti/ → [nEme+ti], which might be the first step to full assimilation, is harmonically bounded by the unchanged candidate.

(39) No harmonic improvement in standard HS

In nEmE+ti	AGREE-ATR	IDENT-ATR
☞ nEmE+ti	*	
nemE+ti	*!*	*
nEme+ti	*	*!

Again, this problem might be resolved by reformulating or replacing AGREE-ATR. However, the LHS solution to the NOCODA problem extends directly to this one.

In the definition of AGREE-ATR that we proposed in (26), the locus of violation is a vowel, and the context is a following vowel that differs in ATR specification. Thus, only a pair of vowels in which the first one is changed by the operation will potentially violate AGREE-ATR in LHS. Again we assume a single operation changing -ATR to +ATR. The portion of the string that is evaluated by AGREE-ATR in LHS is placed in parentheses in the tableau in (40). We return to HG tableaux, since we aim to show that the problem raised in section 4.3 is eliminated under the LHS analysis. But first, we can note that the AGREE-ATR violation that causes the harmonic bounding of the desired intermediate candidate [nEme+ti] in (39) is eliminated by the strict locality of LHS. Therefore, a ranking or weighting of AGREE-ATR above IDENT-ATR picks partial assimilation over the fully faithful candidate.

(40) Harmonic improvement in LHS

Weight	1.5	1	$\mathcal{H}$
In nEmE+ti	AGREE-ATR	IDENT-ATR	
nE(mE+ti)	-1		-1.5
☞ nE(me+ti)		-1	-1

The output of (40) is taken as the input to the next round of change and evaluation.

(41) More harmonic improvement

Weight	1.5	1	$\mathcal{H}$
nEme+ti	AGREE-ATR	IDENT-ATR	
(nEme)+ti	-1		-1.5
☞ (neme)+ti		-1	-1

The procedure now terminates, since there are no more -ATR vowels in the string, and hence no more occasions for change.

We can now see that the problem for HG raised in section 4.3 has been eliminated. With local evaluation, the number of vowels preceding the "trigger" of assimilation is irrelevant. If w(AGREE-ATR) > w(IDENT-ATR), all of the vowels will change, as in this example. If w(IDENT-ATR) > w(AGREE-ATR), then none will. The implausible systems in which n vowels preceding the trigger change and $n + 1$ do not are predicted not to exist.

To show that this account avoids the global OT sour grapes problem, we now consider the case of blocking by **a*. So long as the summed weight of **a* and IDENT-ATR is greater than the weight of AGREE-ATR, harmony is blocked if it would create [a]:

(42) Blocking of harmony

Weight	1	1.5	1	$\mathcal{H}$
In nA+ti	**a*	AGREE-ATR	IDENT-ATR	
☞ (nA+ti)		-1		-1.5
(na+ti)	-1		-1	-2

The sour grapes problem arises when the blocking vowel and the trigger are separated by a string of potential undergoers. Here we provide a case parallel to that in (31). The first step in a LHS derivation is shown in (43):

(43) Blocking of harmony 2

Weight	1	1.5	1	$\mathcal{H}$
In nAmE+ti	**a*	AGREE-ATR	IDENT-ATR	
nA(mE+ti)		-1		-1.5
☞ nA(me+ti)			-1	-1

The second step would be as shown in (42): the /A/ would fail to change to [a]. Because the procedure would cease to produce change, it would terminate.

The key to this LHS solution to the sour grapes problem is that evaluation is strictly local: only the AGREE-ATR violation that is alleviated by the current operation is taken into consideration, while the one that it creates is irrelevant. As mentioned in the paragraph following (31), there are other proposed solutions to this problem. Discussing these accounts would take us too far afield, and we do not intend to present our own approach as a fully worked out theory of vowel harmony, or more generally, of the interaction of operations and constraints in generative linguistics.[8] Instead, we present this discussion of ATR harmony only as an explicit illustration of two points. First, even local constraints have unattested non-local effects when evaluation is global; this was illustrated for HG and OT. Second, local evaluation may resolve these issues for both HG and OT.

In the LHS analysis of ATR harmony presented above, the constraints were either violated or satisfied: there was no need to count the number of violations. This was also true of all the other HS analyses discussed here. Depending on how the Markedness constraints are formulated, the exact degree of violation may differ, but it seems plausible that an upper bound will be placed by the definitions of the constraints, and by the locality convention we have proposed. And as discussed above, HS imposes inherent restrictions on the degree of violation of Faithfulness constraints.

It remains to be better determined how a theory with local evaluation will compare to standard OT in terms of its success in distinguishing possible from impossible languages. Challenges for a theory with local evaluation are apparent non-local interactions, such as the ban on multiple voiced obstruents in Japanese (see fn. 4; see also Rose and Walker 2004 for discussion of long-distance assimilation). It also remains to be better determined how HG and OT fare respectively under local evaluation. As far as we know, the only published example of overgeneration in HG is the one provided by Legendre et al. (2006b), which involves a non-local constraint.

[8] Three main issues would need to be addressed to extend LHS to a broader range of phonological phenomena. First, as mentioned in the text, we have simplified by only considering a single operation at a time; the definition of the interaction between operations and evaluation would need to be elaborated to deal with multiple operations. Second, by limiting the application of constraints to cases in which the locus of violation is affected by the operation, we require reformulation of some constraints. For example, in some languages NOCODA can be satisfied not only by deleting a consonant, but also by adding a vowel, so that the consonant becomes the onset of a syllable. Adding a vowel, however, does not affect the locus of violation in the current formulation. Similarly, AGREE-ATR can be satisfied in some languages by changing not only the first vowel in a disagreeing sequence, but the second one.And finally, cases in which the removal of a phonological element causes a constraint violation would require further elaboration of the theory (Colin Wilson, p.c.).

In the next section, we discuss our proof that with bounded degrees of constraint violation, HG produces only a finite number of languages. This contrasts with global HG, which in our examples, and that of Legendre et al. (2006b), produces an infinite set of possible languages. This can be seen as the most general type of demonstration that HG with local evaluation may be more restrictive than global HG. Before moving to the proof, we make two further comments about the restriction to bounded degrees of violation.

First, we emphasize that boundedness does not necessarily entail a binary distinction between satisfaction and violation. As we mentioned earlier, both positive constraints that reward satisfaction, and scalar constraints that assign a different scores for different degrees of satisfaction or violation can still have an upper bound on degree violation. It may also be that a single operation in LHS can alleviate more than one violation of a single markedness constraint; determining the relationship between single steps in a derivation and the number of possible violations of each constraint would require a more completely elaborated theory. But it does seem in all of these cases of non-binarity of degree of violation, the boundedness assumption may be preserved.

Second, in our proof, we assume the standard OT definitions of candidate sets and languages. These definitions would have to be elaborated for HS and LHS, or for any alternative approach to local evaluation. Because these theories are still in the early stages of development, it is impossible to perform this elaboration now. As some evidence that the proof may generalize, note that the HG/LHS analysis of ATR harmony and coda deletion, and the HG/HS analysis of metathesis, eliminated the instances of infinite sets of languages that obtained under global HG.

5 Finite typologies in HG with bounded constraints

HG can generate an infinity typology in the following sense. The set of potential candidate sets is infinite, since the length of linguistic representations is unbounded. This is true whether the candidate sets themselves are infinite or not (the finiteness of candidate sets depends on assumptions about how candidates are generated). As discussed in the previous section, given an infinite set of candidate sets, it is possible to define constraints that interact with weights in such a way as to allow us to pick infinitely many distinct sets of optima in those candidate sets. For example, in the hypothetical ATR harmony example discussed in 4.3, a distinction between strings of length n and $n+1$ can be made, where n is any integer. There are therefore infinitely many distinctions between optima and sub-optima that this system produces: the set of possible languages is infinite.

This leads naturally to the question of whether there are reasonable limitations we can place on HG that will deliver a finite typology for an infinite set of candidate sets. As noted above, we have identified one such limitation: it is plausible that under local evaluation, all the constraints are functions from candidates into *finite* sets of violation marks. We will use the term violation mark to describe the integer scores the constraints assign because it is familiar from the OT literature, even though constraints in HG may assign integers for both violation and satisfaction. There is an important precedent for the finiteness assumption in the context of OT: Frank and Satta (1998) and Karttunen (1998) show that OT has an attractive finite-state implementation provided that, among other things, the number of violation marks is finite.

In bounding the number of constraint violations, we also bound the number of violation profiles (vectors of violation marks). This is significant because, for the purposes of a given competition, OT grammars cannot distinguish two candidates that have the same violation profile (Samek-Lodovici and Prince 1999; Prince 2002a). The invariance result holds of HG as well, and it brings us much of the way towards the finite typology result.

The remainder of this section explains the result in more detail. A formal proof is given in appendix A.

5.1 Candidates, violation profiles, and optimality

Both OT and HG grammars associate with each input structure I a potentially infinite number of output structures. We call the resulting set of pairs of structures $\mathcal{A}_I$ the candidate set for I. It is the job of the grammar to select, for each input I, the optimal input–output pairs in $\mathcal{A}_I$. In HG, the optimality criterion is (2) above.

Definition (2) makes important use of *violation profiles*: the sequence of violation marks that the constraint set determines for each candidate. More precisely, for a given candidate A and set of n constraints C_n, the violation profile for A is defined as follows:

(44) $\textsc{violations}_{C_n}(A) \stackrel{def}{=} \langle c_i(A), \ldots, c_n(A)\rangle$

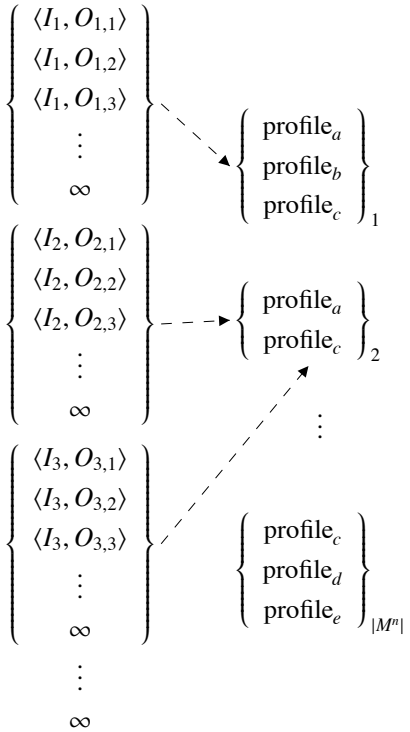
The function $\textsc{violations}_{C_n}$ is assumed to be total for any constraint set C_n and candidate set C_n. Thus, each candidate set $\mathcal{A}_I$ also associates with a particular set of violation profiles — namely, the set containing all and only the violation profiles for some $A \in \mathcal{A}_I$. The function $\textsc{profiles}_{C_n}$ captures this relationship:

(45) Let $\mathcal{A}_I$ be a candidate set for some input I. The function $\textsc{profiles}_{C_n}$ is defined as follows:

$$\textsc{profiles}_{C_n}(\mathcal{A}_I) \stackrel{def}{=} \{v \mid v = \textsc{violations}_{C_n}(A) \text{ for some } A \in \mathcal{A}_I\}$$

We summarize the notions discussed so far in figure 46. For our purposes, the most important aspect of (46) is that, though the set of candidate set is infinite, as is potentially each candidate set it contains, there are only finitely many violation profiles, so the function $\textsc{profiles}_{C_n}$ is many-to-one under all circumstances. The function $\textsc{profiles}_{C_n}$ is a bridge between the hyper-infinite realm of candidate sets and the finite realm of violation profiles.

(46) Illustration of definition (45)

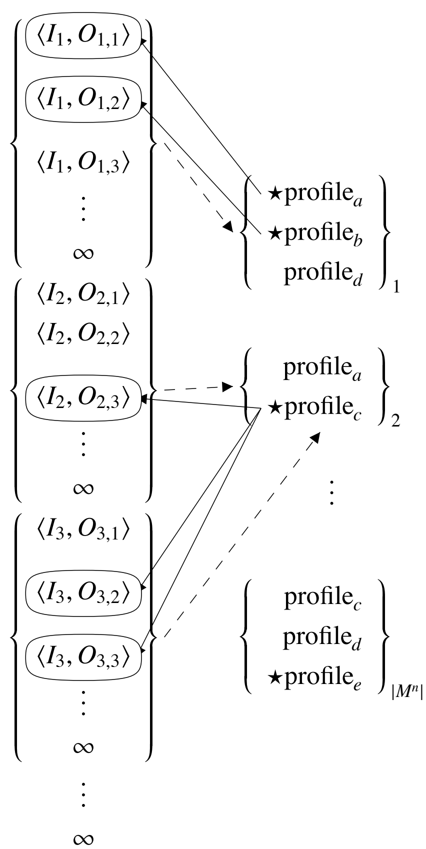


5.2 Language definitions

With $\textsc{profiles}_{C_n}$ taking us to sets of violation profiles, we can call upon an important invariance result of Samek-Lodovici and Prince (1999) and Prince (2002a): if two distinct candidates *A* and *B* are in the same candidate set and determine the same violation profile, then no OT grammar can distinguish them. The same is true in HG. We prove this rigorously (lemma 2), but it is intuitively clear from the definition of harmony, which in effect looks only at a candidate's violations to determine whether it is optimal.

This brings us to the result itself. The language for a given HG is usually defined to be the (infinite) subset of candidates that are optimal according to definition (2). But we obtain an equivalent definition in terms of optimal sets of violation profiles (lemma 3). Figure 47 shows how this selection happens. The dotted lines again represent the $\textsc{profiles}_{C_n}$ relations. We assume that the current weighting W_n selects the starred profiles as optimal. The solid arrows connect these optimal profiles with the candidates who have them in that candidate set. These are circled; they are the optimal candidates, the members of the language.

(47) Languages in terms of violation profiles



Since there are only finitely many ways of selecting distinct subsets of optimal violation profiles, there are only finitely ways of using those optimal profiles to select optimal candidates for inclusion in the language. That is, the predicted typology is finite.

5.3 An upper bound

The finite typologies theorem allows us to state a numerical upper bound on the number of languages predicted for a given HGB. The theorem shows that we can define languages in terms of the sets of violation profiles that associate with candidate sets. There are at most $u = 2^{|M^n|} - 1$ nonempty subsets of violation profiles for a set of marks M and n constraints. We list the elements of this powerset:

$$S_1, S_2, \ldots S_u$$

Each weighting vector selects a nonempty subset of optimal vectors from each S_i. There are $2^{|S_i|} - 1$ ways of making such selections from S_i. So we get the cardinality of each such powerset, subtract 1 from it (for the empty set), and multiply all those values together to get the upperbound on the number of languages.

$$(2^{|S_1|} - 1) \cdot (2^{|S_2|} - 1) \cdot \ldots \cdot (2^{|S_u|} - 1)$$

We emphasize that this is an *upper* bound. In fact, many combinations of winning profiles require inconsistent weighting conditions and cannot form a language.

6 Learning and gradience in HG

Our main goal in this paper is to defend HG against claims that it is insufficiently restrictive to function as a framework for GL. In this section we briefly review and expand upon published arguments for the replacement of OT ranking with HG weighting. These can be divided into two overlapping sets. One broad argument for weighted constraints is that they can cope more successfully with various types of linguistic gradience (Flemming 2001; Keller 2006; Jäger and Rosenbach 2006). The other is that they are compatible with existing well-understood algorithms for learning gradually, learning in noise, and for learning non-categorical patterns (Johnson 2002; Goldwater and Johnson 2003; Jäger 2007; Soderstrom et al. 2006; Wilson 2006a; Pater 2007).

Up to this point, we have adopted a version of HG that like OT is categorical in several ways. The representations it manipulates are category symbols, rather than real-valued continua (cf. Flemming 2001). It makes only a distinction between well-formed mappings (optima) and ill-formed ones (suboptima). It does not distinguish grades of well-formedness (cf. Keller 2006; Legendre et al. 2006a; Hayes and Wilson 2006; Coetzee and Pater 2007). Finally, for each input, the choice of output is invariant across instances of evaluation: the optimum does not vary probabilistically (cf. Goldwater and Johnson 2003; Jäger and Rosenbach 2006; Goldrick and Daland 2007). Here we will consider only the last type of gradience, as it is the one that has been most discussed in the OT literature (e.g., Anttila 1997; Boersma 1998; Boersma and Hayes 2001; Coetzee 2007).

We have also not addressed the question of how constraint weightings are acquired. One of the strengths of OT is that unlike many theories in GL it has associated formal learning algorithms (e.g., Tesar 1995a; Tesar and Smolensky 1998, 2000; Boersma 1998; Boersma and Hayes 2001). This is of course also true of HG, since it was originally framed in terms of a connectionist network (Smolensky and Legendre 2006).

Both variation in the choice of optima for a single word, and issues in learning can be illustrated by the phonology of voicing in Japanese loanwords discussed in section 2. As shown in (10), geminates are optionally devoiced when there is another voiced obstruent in the word (e.g., /doggu/ → [dokku] *or* [doggu]). Our analysis in section 2 only deals with the devoicing outcome; it does not yield the optional lack of devoicing. In terms of learning, Kawahara (2006) emphasizes that the devoicing pattern is emergent: it has entered Japanese in recent borrowings, and there would have

been no evidence in the learners' input for a difference between the forms that show devoicing, and those that do not (e.g., /bobu/ → [bobu] and /webbu/ → [weppu], which lack the combination of a voiced geminate and another voiced obstruent).

We approach these issues by adopting proposals that have been made by Boersma (1998), Boersma and Levelt (2000), and Boersma and Hayes (2001) for OT. The model of grammar assumed by Boersma and colleagues is a hybrid between HG and OT termed stochastic OT. The grammar consists of constraints that are assigned real-valued numbers, which are converted to a corresponding ranked order each time the grammar is used to evaluate a candidate set. To yield variation, the constraint values are perturbed by noise each time the grammar is used. If the values are close enough, their relative order will vary in repeated evaluations of a candidate set, sometimes yielding variation between optima. The associated learning theory consists of an on-line error driven learner that is supplied with correct input–output mappings one at a time. The learner finds the optimum for the input given its own grammar, and if this output fails to match the one in the learning datum, it adjusts the constraint values to favor the mapping in the learning datum over its own *error*.

The learning procedure is very similar to that of the Perceptron learning algorithm (Rosenblatt 1958). For HG, we can adopt the Perceptron's update rule (see Boersma and Pater 2007, Jesney and Tessier 2007 and Pater 2007 for comparisons with the stochastic OT learner; see Jäger 2007 and Soderstrom et al. 2006 for earlier similar proposals in HG and HG-like theories). Given the learner's error, and the correct mapping, each constraint value is updated as in (48).

(48) To the current value, add $n(vE - vC)$

Where n is a constant greater than 0 and less than 1, vE is the violation score for the error, and vC is the violation score for the correct form.

Variation can be handled the same way in HG as in stochastic OT: each time the grammar evaluates a candidate set, the weights are sampled from a normal distribution around the constraint value. We adopt this approach to learning and variation partly because it is so close to one that has been widely used in OT, which allows us to directly compare the HG version to the OT one.

To model the incorporation of English loanwords into a Japanese speaker's phonological system, we start with constraint values that are appropriate for the pattern of voicing in native words: there are no voiced geminates, and only a single voiced obstruent in each word. In (49), we show how the phonology treats English loanwords when a speaker is just beginning to acquire them. Pairs of voiced obstruents, and voiced geminates are both assimilated to the native Japanese pattern by devoicing.

(49) Voicing in native Japanese phonology

a.

	Weight	10	5	$\mathcal{H}$
	In bobu	*2-VOICE	IDENT-VOICE	
	bobu	-1		-10
☞	bopu		-1	-5

b.

	Weight	10	5	$\mathcal{H}$
	In webbu	*VCE-GEM	IDENT-VOICE	
	webbu	-1		-10
☞	weppu		-1	-5

We ran a learning simulation in Praat (Boersma and Weenink 2007), which implements an error-driven learner with HG evaluation, and the update rule in (48). The learner was supplied with the mappings /bobu/ → [bobu], /webbu/ → [webbu] and /doggu/ → [doggu] with equal probability. The initial weightings were as in (49), and the learning rate (n in (48)) was set at 0.1. After 150 pieces of learning data, the weightings were as shown in (50). In these tableaux, the percentages indicate the frequency with which different mappings emerge as optimal when the weightings are sampled from a normal distribution with a standard deviation of 2 around the weighting value. With these weights, /doggu/ shows much more variation than the others, which almost always surface with voicing intact.

(50) Emergent variable Japanese loanword devoicing

a.

	Weight	11.4	4.7	$\mathcal{H}$
	In bobu	IDENT-VOICE	*2-VOICE	
99%	bobu		-1	-4.7
1%	bopu	-1		-11.4

b.

	Weight	11.4	4.4	$\mathcal{H}$
	In webbu	IDENT-VOICE	*VCE-GEM	
99%	webbu		-1	-4.7
1%	weppu	-1		-11.4

c.

	Weight	11.4	4.4	4.7	$\mathcal{H}$
	In doggu	IDENT-VOICE	*VCE-GEM	*2-VOICE	
74%	doggu		-1	-1	-9.1
26%	dokku	-1			-11.4

This learning simulation is a gross abstraction from the actual process of loanword incorporation, and results that more closely match the available Japanese data could doubtless be obtained by making it more realistic in various ways, and also perhaps by changing aspects of the model of grammar or learning. It nonetheless serves to illustrate three points.

First, the variable pattern of devoicing in Japanese loanwords emerges directly from the basic assumptions that learning is gradual, and that evaluation is noisy. While further elaboration is needed to explain the apparent stability of this pattern in current Japanese, the fact that it is a predicted stage of learning should render this elaboration tractable.

Second, HG is compatible with the simple Perceptron learning algorithm, and many other more elaborate connectionist and statistical learners. Perceptron is attractive not only for its simplicity, but also for the fact that it provably converges on a correct set of weights. Collins (2002) extends the Perceptron convergence proof to an application in part-of-speech tagging, and Boersma and Pater (2007) show how Collins' proof can be applied to the case of learning in HG.

Third, HG can be straightforwardly elaborated to yield variation. Standard OT is purely categorical, and the standard OT learning algorithm fails to converge when supplied with variable data (Tesar and Smolensky 1998). With the exception of Jarosz (2006), accounts of variation in OT either have no associated learning algorithm (e.g., Anttila 1997, Coetzee 2007), or have one that relies on incorporating HG-like weighting values and is demonstrably nonconvergent (see Pater 2007 on the learner for stochastic OT in Boersma 1998; Boersma and Hayes 2001). The Perceptron+noisy HG combination we used in our simulation does not have convergence proof, but testing shows that it is robust (Boersma and Pater 2007). The log-linear models of grammar in Johnson 2002; Goldwater and Johnson 2003; Jäger 2007; Wilson 2006a; Jäger and Rosenbach 2006 calculate the probability of candidate mappings as proportional to the exponential of their harmony. They have associated provably convergent learning algorithms,

but they yield patterns of variation somewhat different from the other models discussed here (in particular, they can give harmonically bounded candidates a portion of the probability mass; Jäger and Rosenbach 2006; **?**).

The main strength of Tesar and Smolensky's OT learning algorithm is that it is detects infeasibility: given a set of optima with associated candidate sets and violation profiles, it will return a ranking, else indicate that none exists. This is particularly useful for constructing analyses of categorical patterns, and for studying the typological predictions of constraint sets (see Hayes et al. 2003). As we show in a related paper, infeasibility detection is also possible in categorical HG, by applying the simplex method of Linear Programming (Bhatt et al. 2007). We hope that the publicly available implementation of this application (Potts et al. 2007) will facilitate further investigation of HG as a framework for linguistic analysis and typological study.

7 Conclusions

HG remains relatively unexplored in generative linguistics. Given our current understanding of the theory, and its relationship to OT, it appears to be a viable framework for research on the formal properties of natural languages. Its core property of optimization is both powerful and restrictive. We believe that a key aspect of the further development of this framework will be the elaboration of the notion of local iterative evaluation: the previously identified problems with global evaluation in OT are amplified in HG. This elaboration will require reassessment of basic assumptions in current linguistic theory. Most research in OT has operated under the assumption of global evaluation, while much generative research outside of OT in both phonology and syntax has mostly developed the iterative change component (that of *rules* or *transformations*), and has paid less attention to constraints. We see much promise in bringing these lines of research together, as in McCarthy's (2006 *et seq.*) recent work in phonology. Finally, from the perspective of the cognitive architecture proposed by Smolensky and Legendre (2006), the replacement of OT ranking with HG weighting may contribute to the resolution of a longstanding problem: that there is no known method for making a connectionist network behave as an OT strict domination hierarchy (Prince and Smolensky 2004: 236; Legendre et al. 2006b: 347).

A Proof of finite typologies in HGB

This appendix presents a formal proof of the finite typologies result described in section 5. We repeat some definitions given in the main text in order to make the proof as self-contained as possible.

A.1 Harmonic grammars with bounded constraints

Definition 1 (HGBs)**.** *A harmonic grammar with bounded constraints (HGB) is a tuple* $(\Sigma, \textsc{Gen}, M, C_n, W_n)$*, where*

a. Σ *is a finite alphabet;*

b. Gen *is a function from* Σ^* *to* $\wp(\Sigma^*)$*;*

c. M *is a finite set of integers;*

d. $C_n = \{c_1 \dots c_n\}$ *is a set of n distinct total functions from* $\Sigma^* \times \Sigma^*$ *into* M*; and*

e. $W_n = \langle w_1 \dots w_n \rangle$ *is a vector of real numbers.*

A.2 Preliminary definitions

We first define a series of supplementary notions. Throughout, we assume an HGB $\mathbf{H} = (\Sigma, \textsc{Gen}, M, C_n, W_n)$.

Definition 2 (Candidates)**.** *A pair $\langle I, O\rangle$ is an $\mathbf{H}$-candidate iff $I \in \Sigma^*$ and $O \in \textsc{Gen}(I)$.*

Definition 3 (Candidate sets)**.** *The candidate set $\mathcal{A}_I$ for an input I is the set of candidates $\langle I, O\rangle$ such that $O \in \textsc{Gen}(I)$.*

Definition 4 (Violation profiles)**.** *The function $\textsc{violations}_{C_n}$ maps candidates to their violation profiles for the constraint set $C_n = \langle c_1 \dots c_n\rangle$:*

$$\textsc{violations}_{C_n}(A) \stackrel{def}{=} \langle c_1(A) \dots c_n(A)\rangle$$

Definition 5. *The set of all violation profiles for the set of marks M and the set of constraints C_n is*

$$M^n \stackrel{def}{=} M_1 \times \dots \times M_n$$

We use $|M^n|$ for the number of violation profiles.

Fact 1. *$|M^n|$ is finite. (This follows from the fact C_n and M are both finite.)*

Definition 6 (Weighted violation counts)**.** *The weighted violation count for a candidate A given constraint set $C_n = \langle c_1 \dots c_n\rangle$ and weights $W_n = \langle w_1 \dots w_n\rangle$ is defined as follows:*

$$\mathcal{H}_{W_n,C_n}(A) \stackrel{def}{=} \sum_{j=1}^{n} w_j(c_j(A))$$

Definition 7 (Optimality; Smolensky and Legendre 2006)**.** *A candidate $A = \langle I, O\rangle$ is optimal in candidate set $\mathcal{A}_I$ iff*

$$\mathcal{H}_{W_n,C_n}(A) \geqslant \mathcal{H}_{W_n,C_n}(A') \quad \textit{for all } A' \in \mathcal{A}_I$$

We write $\textsc{optimal}_{W_n,C_n}(A, \mathcal{A})$ for this.

A.3 candidate sets to sets of violation profiles

The purpose of this section is to show that we can reason about candidates entirely in terms of their violation profiles. This is the major step towards our finiteness result, since it moves us from the infinite domain of candidates into the finite domain of violation profiles.

Our first lemma simply shows that a candidate's $\mathcal{H}_{W_n,C_n}$ value is the same as the weighted sum of its violation marks.

Lemma 1. *Let $(\Sigma, \textsc{Gen}, M, C_n, W_n)$ be a harmonic grammar with bounded constraints, and let A be a candidate. Then $\mathcal{H}_{W_n,C_n}(A)$ is equivalent to the sum of all the violations in $\textsc{violations}_{C_n}(A)$.*

Proof. Let $\textsc{violations}_{C_n}(A)_i$ be the ith member of $\textsc{violations}_{C_n}(A)$. Then we just need to show that

$$\sum_{j=1}^{n} w_j(c_j(A)) = \sum_{j=1}^{n} w_j \textsc{violations}_{C_n}(A)_j$$

This is immediate from the fact that $c_j(A)$ just is $\textsc{violations}_{C_n}(A)_j$, for all $j = 1 \dots n$, by definition 4. □

We now define what it means to be an optimal profile. The definition is parallel to definition 7, but it makes no mention of constraints or candidates.

Definition 8 (Optimality of violation profiles)**.** *Given an HGB* $(\Sigma, \textsc{Gen}, M, C_n, W_n)$ *and a set of violation profiles* $V \subseteq M^n$*, a profile* $v \in V$ *is optimal iff*

$$\sum_{j=1}^{n} w_j v_j \geqslant \sum_{j=1}^{n} w_j v'_j \quad \textit{for all } v' \in V$$

We write $\textsc{vp-optimal}_{W_n}(v, V)$ *for this.*

We now define a mapping from candidate sets to sets of violation profiles:

Definition 9 (Candidate sets to sets of violation profiles)**.** *Let* $(\Sigma, \textsc{Gen}, M, C_n, W_n)$ *be a harmonic grammar with bounded constraints, and let* $\mathcal{A}_I$ *be a candidate set for some* $I \in \Sigma^*$*. The function* $\textsc{profiles}_{C_n}$ *is defined as follows:*

$$\textsc{profiles}_{C_n}(\mathcal{A}_I) \stackrel{def}{=} \{v \mid v = \textsc{violations}_{C_n}(A) \textit{ for some } A \in \mathcal{A}_I\}$$

Figure 46 (section 5) provides a sense for how $\textsc{profiles}_{C_n}$ works. The arrows represent the action of $\textsc{profiles}_{C_n}$ on each candidate set, mapping it to its associated set of violation profiles (as determined by the profiles of the candidates in that candidate set).

Fact 2. *The function* $\textsc{profiles}_{C_n}$ *maps candidate sets to subsets of the set of violation profiles. It is total and many-to-one. It is total because candidate sets consist of candidates, and each constraint is a total function on candidates. It is many-to-one because the set of all candidate sets is infinite, since there is one for every* $I \in \Sigma^*$*. But, by fact 1, the set of all violation profiles is finite, and hence so is its powerset.*

We're now ready to prove the central lemma showing that we can reason about optimality entirely in terms of sets of violation profiles.

Lemma 2. *Let* $(\Sigma, \textsc{Gen}, M, C_n, W_n)$ *be a harmonic grammar with bounded constraints, and* $A = \langle I, O \rangle$ *be a candidate in* $\mathcal{A}_I$*. Then*

$$\textsc{optimal}_{W_n,C_n}(A, \mathcal{A}_I) \quad \Longleftrightarrow \quad \textsc{vp-optimal}_{W_n}(\textsc{violations}_{C_n}(A), \textsc{profiles}_{C_n}(\mathcal{A}_I))$$

That is, a candiate is optimal iff its violation profile is optimal.

Proof. We first prove the left-to-right implication, then the right-to-left one.

[⇒] Assume $\textsc{optimal}_{W_n,C_n}(A, \mathcal{A}_I)$.

Suppose it is false that $\textsc{vp-optimal}_{W_n}(\textsc{violations}_{C_n}(A), \textsc{profiles}_{C_n}(\mathcal{A}_I))$. We show that this leads to a contradiction.

Let v be a profile such that $\textsc{vp-optimal}_{W_n}(v, \textsc{profiles}_{C_n}(\mathcal{A}_I))$. Then, by definition 8, it holds that

$$\sum_{j=1}^{n} w_j v_j > \sum_{j=1}^{n} w_j \textsc{violations}_{C_n}(A)_j \tag{51}$$

Let B be a candidate in $\mathcal{A}_I$ such that $\textsc{violations}_{C_n}(B) = v$. By lemma 1 and equation 51, it therefore holds that

$$\mathcal{H}_{W_n,C_n}(B) > \mathcal{H}_{W_n,C_n}(A)$$

But then it is false that $\textsc{optimal}_{W_n,C_n}(A, \mathcal{A}_I)$, by definition 7. This contradicts our initial assumption.

[⇐] Assume $\text{VP-OPTIMAL}_{W_n}(\text{VIOLATIONS}_{C_n}(A), \text{PROFILES}_{C_n}(\mathcal{A}_I))$.

Suppose it is false that $\text{OPTIMAL}_{W_n,C_n}(A, \mathcal{A}_I)$. We show that this leads to a contradiction.

Let B be the candidate such that $\text{OPTIMAL}_{W_n,C_n}(B, \mathcal{A}_I)$. Then

$$\mathcal{H}_{W_n,C_n}(B) > \mathcal{H}_{W_n,C_n}(A)$$

By lemma 1, this means that

$$\sum_{j=1}^{n} w_j \text{VIOLATIONS}_{C_n}(B)_j > \sum_{j=1}^{n} w_j \text{VIOLATIONS}_{C_n}(A)_j$$

But this contradicts the assumption that $\text{VP-OPTIMAL}_{W_n}(\text{VIOLATIONS}_{C_n}(A), \text{PROFILES}_{C_n}(\mathcal{A}_I))$. □

A.3.0.1 The main result of the above Lemma 2 essentially says that we need not calculate optimality in terms of comptetitor classes. Rather, we can calculate it in terms of the set of violation profiles that each candidate set determines (by definition 9). This is an important gain, because while candidate sets can be infinite, their associated sets of violation profiles are finite.

A.4 Typology

We now move to the typological predictions. We first define languages in terms of optimal candidates, then in terms of optimal profiles. Lemma 3 shows that the definitions are equivalent. This leads fairly directly to our theorem.

Definition 10 (Languages)**.** *The language generated by* $\mathbf{H} = (\Sigma, \text{GEN}, M, C_n, W_n)$ *is a set*

$$\mathcal{L}_{\mathbf{H}} \stackrel{def}{=} \{\langle I, O\rangle \mid \text{OPTIMAL}_{W_n,C_n}(\langle I, O\rangle, \mathcal{A}_I), \textit{ where } \mathcal{A}_I \textit{ is the candidate set for } \langle I, O\rangle\}$$

This is equivalent to the union of the set of all optimal candidates. We present it in this way to emphasize that we effectively select the winning candidates from each candidate set to form the language.

Definition 11. *Let* $\mathbf{H} = (\Sigma, \text{GEN}, M, C_n, W_n)$ *be a harmonic grammar with bounded constraints. Let* C *be the set of all candidate sets:*

$$C \stackrel{def}{=} \{\mathcal{A}_I \mid I \in \Sigma^*\}$$

We define the corresponding set of sets of violation profiles:

$$\mathcal{V}_{\mathbf{H}} \stackrel{def}{=} \{V \mid \text{PROFILES}_{C_n}(\mathcal{A}_I) = V \textit{ for some } \mathcal{A}_I \in C\}$$

Fact 3. $\mathcal{V}_{\mathbf{H}}$ *is defined independently of the weights.* $\mathcal{V}_{\mathbf{H}}$ *is finite, each of its members is finite, and a subset of its members are optimal. Hence, there are only finitely many ways of selecting winning profiles.*

Definition 12 (Inverse of PROFILES_{C_n})**.** *We define the set-valued inverse of* PROFILES_{C_n}*:*

$$\text{PROFILES}_{C_n}^{-1}(V) \stackrel{def}{=} \{\mathcal{A} \mid \text{PROFILES}_{C_n}(\mathcal{A}) = V\}$$

Fact 4. $\text{PROFILES}_{C_n}^{-1}$ *partitions the set of candidate sets into equivalences classes based on values for* PROFILES_{C_n}*.*

Definition 13 (Languages via violation profiles)**.** *Let* $(\Sigma, \textsc{Gen}, M, C_n, W_n)$ *be a harmonic grammar with bounded constraints, and let* C *and* $\mathcal{V}_{\mathbf{H}}$ *be as in definition 11. The language generated by* $\mathcal{V}_{\mathbf{H}}$ *given* W_n *is defined as follows:*

$$\mathcal{L}_{\mathcal{V}_{\mathbf{H}},W_n} \stackrel{def}{=} \{\langle I, O\rangle \mid \exists V \in \mathcal{V}_{\mathbf{H}} : \mathcal{A}_I \in \textsc{profiles}_{C_n}^{-1}(V) \textit{ and } \textsc{vp-optimal}_{W_n}(\textsc{violations}_{C_n}(\langle I, O\rangle), V)\}$$

Figure 47 (section 5) explains how this kind of language selection works. We now show that it selects the same languages as definition 10 does.

Lemma 3. *Let* $(\Sigma, \textsc{Gen}, M, C_n, W_n)$ *be a harmonic grammar with bounded constraints. Then* $\mathcal{L}_{\mathbf{H}} = \mathcal{L}_{\mathcal{V}_{\mathbf{H}}}$*. (That is, definitions 10 and 13 are equivalent.)*

Proof. **Case 1.** Let $A = \langle I, O\rangle$ be such that $A \notin \mathcal{L}_{\mathcal{V}_{\mathbf{H}}}$. We show that this entails that $A \notin \mathcal{L}_{\mathbf{H}}$.

Since $\textsc{profiles}_{C_n}^{-1}$ partitions the space of candidate sets, it is guaranteed that there is a V such that $\mathcal{A}_I \in \textsc{profiles}_{C_n}^{-1}(V)$, where $\mathcal{A}_I$ is the candidate set for A. Hence, if $A \notin \mathcal{L}_{\mathcal{V}_{\mathbf{H}}}$, then it is false that $\textsc{vp-optimal}_{W_n}(\textsc{violations}_{C_n}(A), \textsc{profiles}_{C_n}(\mathcal{A}_I))$. By lemma 2, this entails that it is false that $\textsc{optimal}_{W_n,C_n}(A, \mathcal{A}_I)$. By definition 10, this means that $A \notin \mathcal{L}_{\mathbf{H}}$.

Case 2. Let $A = \langle I, O\rangle$ be such that $A \notin \mathcal{L}_{\mathbf{H}}$. We show that this entails that $A \notin \mathcal{L}_{\mathcal{V}_{\mathbf{H}}}$.

If $A \notin \mathcal{L}_{\mathbf{H}}$, then it is false that $\textsc{optimal}_{W_n,C_n}(A, \mathcal{A}_I)$. Thus, it is false that $\textsc{vp-optimal}_{W_n}(\textsc{violations}_{C_n}(A), \textsc{profiles}_{C_n}(\mathcal{A}_I))$, by lemma 2. Thus, by definition 13, $A \notin \mathcal{L}_{\mathcal{V}_{\mathbf{H}}}$. □

Definition 14 (Typology)**.** *The typology for* Σ*,* M*,* Gen*, and* C_n *is the set of all languages generated by* $\mathbf{H} = (\Sigma, \textsc{Gen}, M, C_n, W_n)$ *for some weighting vector* W_n*.*

Theorem (Finite HGB typologies). *The number of distinct languages in the typology for* Σ*,* M*,* Gen*, and* C_n *is finite.*

Proof. Lemma 3 ensures that we can define languages in terms of sets of optimal violation profiles or in terms of optimal candidates, with equivalent results in both cases. Fact 1 says that there are only finitely many violation profiles. Therefore, $\mathcal{V}$ in definition 11 is always a finite set of finite sets. Fact 3 says that there are only finitely many ways of selecting $\textsc{vp-optimal}_{W_n}$ profiles from each one, and hence only finitely many sets of candidates can be defined by definition 13 (despite the fact that there are infinitely many W_ns). Hence, the typology for Σ, M, Gen, and C_n is finite. □

References

Alderete, J. (1997). Dissimilation as local conjunction. In Kusumoto, K. and Kusumoto, K., editors, *Proceedings of the North East Linguistic Society 27*, pages 17–32. GLSA Publications, Amherst, Mass.

Anttila, A. (1997). Deriving variation from grammar. In Hinskens, F., van Hout, R., and Wetzels, W. L., editors, *Variation, Change, and Phonological Theory*, pages 35–68. John Benjamins, Amsterdam.

Bakovic, E. (2000). *Harmony, Dominance, and Control*. PhD thesis, Rutgers University, New Brunswick, NJ.

Bhatt, R., Pater, J., and Potts, C. (2007). Harmonic Grammar with Linear Programming. Ms., UMass Amherst.

Bíró, T. (2003). Quadratic alignment constraints and finite state optimality theory. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing*, pages 119–126, Budapest, Hungary. Available at http://roa.rutgers.edu/.

Bíró, T. (2006). *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*. PhD thesis, University of Groningen.

Boersma, P. (1998). *Functional Phonology: Formalizing the Interaction Between Articulatory and Perceptual Drives*. Holland Academic Graphics, The Hague.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32:45–86. Available on Rutgers Optimality Archive, http://ruccs.rutgers.edu/roa.html.

Boersma, P. and Levelt, C. C. (2000). Gradual constraint-ranking learning algorithm predicts acquisition order. In Clark, E. V. and Clark, E. V., editors, *The Proceedings of the Thirtieth Annual Child Language Research Forum*. CSLI Publications, Stanford, CA. Available on Rutgers Optimality Archive, http://ruccs.rutgers.edu/roa.html.

Boersma, P. and Pater, J. (2007). Testing gradual learning agorithms. Ms, University of Amsterdam and UMass Amherst.

Boersma, P. and Weenink, D. (2007). Praat: doing phonetics by computer (Version 4.6) [Computer program]. Retrieved May 16, 2007 from http://www.praat.org/. Developed at the Institute of Phonetic Sciences, University of Amsterdam.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. (1985). *The Logical Structure of Linguistic Theory*. University of Chicago Press, Chicago.

Chomsky, N. (2000). Minimalist inquiries: The framework. In Martin, R., Michaels, D., and Uriagereka, J., editors, *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, pages 89–156. MIT Press.

Chomsky, N. (2001). Beyond explanatory adequacy. Occasional Papers in Linguistics.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.

Chomsky, N. and Lasnik, H. (1977). Filters and control. *Linguistic Inquiry*, 8:425–504.

Coetzee, A. (2007). Variation as accessing 'non-optimal' candidates: A rank-ordering model of eval. *Phonology*. To appear.

Coetzee, A. and Pater, J. (2007). Weighted constraints and gradient phonotactics in Muna and Arabic. Ms, University of Michigan and UMass Amherst.

Collins, M. (2002). Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Dresher, E. (1994). Acquiring stress systems. In Ristad, E. S., editor, *Language Computations*, number 17 in DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pages 71–92. AMS, Providence, Rhode Island.

Eisner, J. (1997). Efficient generation in Primitive Optimality Theory.

Eisner, J. (1998). FootForm decomposed: Using primitive constraints in OT. In Bruening, B., editor, *Proceedings of SCIL VIII*, number 31 in MIT Working Papers in Linguistics, pages 115–143, Cambridge, MA.

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, 18(1):7–44.

Frank, R. and Satta, G. (1998). Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics*, 24:307–315.

Goldrick, M. and Daland, R. (2007). Linking grammatical principles with experimental speech production data: Insights from harmonic grammar networks. In *Paper presented at Experimental Approaches to Optimality Theory.*, Ann Arbor, MI. Slides available at http://ling.northwestern.edu/ goldrick/ExpOT.pdf.

Goldsmith, J. (1990). *Autosegmental and Metrical Phonology*. Blackwell, Oxford and Cambridge, Mass.

Goldsmith, J. (1993). Harmonic phonology. In Goldsmith, J., editor, *The Last Phonological Rule*, pages 21–60. University of Chicago Press, Chicago.

Goldsmith, J. (1994). A dynamic computational theory of accent systems. In Cole, J. and Kisseberth, C., editors, *Perspectives in Phonology*, pages 1–28. CSLI, Stanford.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In Spenader, J., Eriksson, A., and Dahl, Ö., editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120. Stockholm University, Stockholm.

Gordon, M. (2002). A factorial typology of quantity-insensitive stress. *Natural Language and Linguistic Theory*, 20:491–552.

Gupta, P. and Touretzky, D. S. (1994). Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science*, 18(1):1–50.

Hale, K. (1973). Deep-surface canonical disparities in relation to analysis and change: An Australian example. In Sebeok, T., editor, *Current Trends in Linguistics*, volume 9: Diachronic, Areal and Typological Linguistics, pages 401–458. Mouton, The Hague.

Halle, M. and Vergnaud, J.-R. (1987). *An Essay on Stress*. MIT Press, Cambridge, Mass.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. The University of Chicago Press, Chicago.

Hayes, B., Tesar, B., and Zuraw, K. (2003). OTSoft 2.1. Software package developed at UCLA.

Hayes, B. and Wilson, C. (2006). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*. To appear.

Ito, J. and Mester, A. (1986). The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry*, 17:49–73.

Ito, J. and Mester, A. (2003). *Japanese Morphophonemics: Markedness and Word Structure*. MIT Press Linguistic Inquiry Monograph Series 41, Cambridge, Massachusetts.

Jäger, G. (2007). Maximum entropy models and Stochastic Optimality Theory. In Grimshaw, J., Maling, J., Manning, C., Simpson, J., and Zaenen, A., editors, *Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan*. CSLI, Stanford, CA.

Jäger, G. and Rosenbach, A. (2006). The winner takes it all - almost. cumulativity in grammatical variation. *Linguistics*, 44:937–971.

Jarosz, G. (2006). *Rich lexicons and restrictive grammars - Maximum likelihood learning in Optimality Theory*. PhD thesis, Johns Hopkins University, Baltimore, Maryland. ROA-884.

Jesney, K. and Tessier, A.-M. (2007). Re-evaluating learning biases in harmonic grammar. *University of Massachusetts Working Papers in Linguistics*. To appear.

Johnson, D. E. and Postal, P. M. (1980). *Arc Pair Grammar*. Princeton University Press, Princeton, NJ.

Johnson, M. (2002). Optimality-theoretic lexical functional grammar. In Stevenson, S. and Merlo, P., editors, *The Lexical Basis of Syntactic Processing: Formal, Computational and Experimental Issues*, pages 59–73. John Benjamins.

Joshi, A. K., Levy, L., and Takahashi, M. (1975). Tree adjunct grammars. *Journal of the Computer and System Sciences*, 10(1):136–163.

Kager, R. (2001). Rhythmic directionality by positional licensing. In *Handout of a paper presented at the HILP conference, Potsdam 2001.*, University of Potsdam. Available on the Rutgers Optimality Archive, ROA 514, http://roa.rutgers.edu.

Kaplan, R. M. and Bresnan, J. (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge, MA.

Karttunen, L. (1998). The proper treatment of optimality in computational phonology. Ms., Xerox Research Centre Europe, Meylan, France. ROA 258-0498.

Kawahara, S. (2006). A faithfulness ranking projected from a perceptibility scale: The case of [+voice] in japanese. *Language*, 82(3):536–574.

Keller, F. (2006). Linear optimality theory as a model of gradience in grammar. In Fanselow, G., Féry, C., Vogel, R., and Schlesewsky, M., editors, *Gradience in Grammar: Generative Perspectives*. Oxford University Press, Oxford.

Legendre, G., Miyata, Y., and Smolensky, P. (2006a). The interaction of syntax and semantics: A Harmonic Grammar account of split intransitivity. In Smolensky and Legendre (2006), pages 417–452.

Legendre, G., Sorace, A., and Smolensky, P. (2006b). The Optimality Theory–Harmonic Grammar connection. In Smolensky and Legendre (2006), pages 339–402.

Lombardi, L. (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory*, 17:267–302.

Lubowicz, A. (2005). Locality of Conjunction. In *Proceedings of WCCFL 24*. Cascadilla, Somerville, MA.

McCarthy, J. (2007). *Hidden Generalizations: Phonological Opacity in Optimality Theory*. Advances in Optimality Theory. Equinox, London, UK.

McCarthy, J. J. (1999). Sympathy and phonological opacity. *Phonology*, 16:331–399.

McCarthy, J. J. (2002). *A Thematic Guide to Optimality Theory*. Cambridge University Press, Cambridge.

McCarthy, J. J. (2003a). Comparative Markedness. *Theoretical Linguistics*, 29(29):1–51.

McCarthy, J. J. (2003b). OT constraints are categorical. *Phonology*, 20(1):75–138.

McCarthy, J. J. (2004). Headed spans in autosegmental spreading. Available as ROA-685 on the Rutgers Optimality Archive, http://roa.rutgers.edu.

McCarthy, J. J. (2006). Restraint of analysis. In *Wondering at the Natural Fecundity of Things: Essays in Honor of Alan Prince*, chapter 10. Linguistics Research Center.

McCarthy, J. J. and Prince, A. (1986). Prosodic Morphology 1986. Excerpts appear in John Goldsmith, ed., Essential Readings in Phonology. Oxford: Blackwell. Pp. 102-136, 1999.

McCarthy, J. J. and Prince, A. (1993). Generalized Alignment. In Booij, G. and van Marle, J., editors, *Yearbook of Morphology*, pages 79–153. Kluwer, Dordrecht. Excerpts appear in John Goldsmith, ed., Essential Readings in Phonology. Oxford: Blackwell. Pp. 102?136, 1999.

McCarthy, J. J. and Prince, A. (1999). Faithfulness and identity in Prosodic Morphology. In Kager, R., van der Hulst, H., and Zonneveld, W., editors, *The Prosody-Morphology Interface*, pages 218–309. Cambridge University Press, Cambridge.

McCawley, J. D. (1968). Concerning the base component of a transformational grammar. *Foundations of Language*, 4(1):55–88.

Nishimura, K. (2003). Lyman's law in loanwords. Ms, University of Tokyo.

Paradis, C. (1988). On constraints and repair strategies. *The Linguistic Review*, 6:71–97.

Pater, J. (2007). Gradual learning and convergence. *Linguistic Inquiry*. To appear.

Pinker, S. and Prince, A. (1988). On Language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28:73–193.

Potts, C., Becker, M., Bhatt, R., and Pater, J. (2007). HaLP: Harmonic grammar with linear programming, version 2. Software available online at http://web.linguist.umass.edu/~halp/.

Prince, A. (1993). In defense of the number i: Anatomy of a linear dynamical model of linguistic generalizations.

Prince, A. (2002a). Anything goes. In Honma, T., Okazaki, M., Tabata, T., and ichi Tanaka, S., editors, *New Century of Phonology and Phonological Theory*, pages 66–90. Kaitakusha, Tokyo. ROA-536.

Prince, A. (2002b). Arguing optimality. In Coetzee, A., Carpenter, A., and de Lacy, P., editors, *University of Massachusetts Occasional Papers in Linguistics: Papers in Optimality Theory II*. GLSA, UMass Amherst. Available on the Rugers Optimality Archive, ROA 562.

Prince, A. and Smolensky, P. (1993/2004). Optimality Theory: Constraint interaction in generative grammar. RuCCS Technical Report 2, Rutgers University, Piscateway, NJ: Rutgers University Center for Cognitive Science. Revised version published 2004 by Blackwell. Page references to the 2004 version.

Prince, A. and Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, 275:1604–1610.

Pullum, G. K. and Scholz, B. C. (2001). On the distinction between model-theoretic and generative–enumerative syntactic frameworks. In de Groote, P., Morrill, G., and Retoré, C., editors, *Logical Aspects of Computational Linguistics: 4th International Conference, LACL 2001*, pages 17–43. Springer.

Riggle, J. (2004). *Generation, Recognition and Learning in Finite State Optimality Theory*. PhD thesis, UCLA.

Riggle, J. and Wilson, C. (2005). Local optionality. In *Proceedings of NELS 35*. GLSA, Cascadilla Press. Available at http://www.linguistics.ucla.edu/people/wilson/papers.html.

Rose, S. and Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, 80(3):475–532.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408.

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tense of English verbs. In McClelland, J. L., Rumelhart, D. E., Group, t. P. R., McClelland, J. L., Rumelhart, D. E., and Group, t. P. R., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, pages 216–271. MIT Press/Bradford Books, Cambridge, Mass.

Samek-Lodovici, V. and Prince, A. (1999). Optima. RuCCS Technical Report no. 57, Rutgers University, Piscateway, NJ: Rutgers University Center for Cognitive Science. ROA 363-1199.

Smolensky, P. (2006). Optimality in phonology II: Harmonic completeness, local constraint conjunction, and feature domain markedness. In Smolensky and Legendre (2006), pages 27–160.

Smolensky, P. and Legendre, G. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. MIT Press, Cambridge, MA.

Soderstrom, M., Mathis, D. W., and Smolensky., P. (2006). Abstract genomic encoding of universal grammar in optimality theory. In Smolensky, P. and Legendre, G., editors, *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar. Volume 2: Linguistic and Philosophical Implications.*, pages 403–471. MIT Press.

Steriade, D. (1999). Phonetics in phonology: the case of laryngeal neutralization. In Gordon, M., editor, *Papers in Phonology 3 (UCLA Working Papers in Linguistics 2)*, pages 25–145. Department of Linguistics, University of California, Los Angeles.

Suzuki, K. (1998). *A Typological Investigation of Dissimilation*. PhD thesis, University of Arizona, Tucson, AZ.

Tesar, B. (1995a). *Computational Optimality Theory*. PhD thesis, University of Colorado at Boulder. ROA 90-0000.

Tesar, B. (1995b). Computing Optimal Forms in Optimality Theory: Basic Syllabification. Technical Report.

Tesar, B. and Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, 29:229–268.

Tesar, B. and Smolensky, P. (2000). *Learnability in Optimality Theory*. MIT Press, Cambridge, Mass.

Wilson, C. (2003). Analyzing unbounded spreading with constraints: marks, targets and derivations. *Unpublished ms, UCLA*.

Wilson, C. (2006a). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5):945–982.

Wilson, C. (2006b). Unbounded spreading is myopic. Available at http://www.linguistics.ucla.edu/people/wilson/talks.html.