

January 2009

# Weighted Constraints in Generative Linguistics

Joe Pater  
pater@linguist.umass.edu

Follow this and additional works at: [https://scholarworks.umass.edu/linguist\\_faculty\\_pubs](https://scholarworks.umass.edu/linguist_faculty_pubs)



Part of the [Linguistics Commons](#)

---

## Recommended Citation

Pater, Joe, "Weighted Constraints in Generative Linguistics" (2009). *Cognitive Science*. 173.  
10.1111/j.1551-6709.2009.01047.x

This Article is brought to you for free and open access by the Linguistics at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Linguistics Department Faculty Publication Series by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).

**University of Massachusetts Amherst**

---

**From the Selected Works of Joe Pater**

---

2009

# Weighted Constraints in Generative Linguistics

Joe Pater



Available at: [https://works.bepress.com/joe\\_pater/18/](https://works.bepress.com/joe_pater/18/)

# Weighted Constraints in Generative Linguistics

Joe Pater

*Department of Linguistics, University of Massachusetts, Amherst*

Received 22 July 2007; received in revised form 15 March 2009; accepted 24 March 2009

---

## Abstract

Harmonic Grammar (HG) and Optimality Theory (OT) are closely related formal frameworks for the study of language. In both, the structure of a given language is determined by the relative strengths of a set of constraints. They differ in how these strengths are represented: as numerical weights (HG) or as ranks (OT). Weighted constraints have advantages for the construction of accounts of language learning and other cognitive processes, partly because they allow for the adaptation of connectionist and statistical models. HG has been little studied in generative linguistics, however, largely due to influential claims that weighted constraints make incorrect predictions about the typology of natural languages, predictions that are not shared by the more popular OT. This paper makes the case that HG is in fact a promising framework for typological research, and reviews and extends the existing arguments for weighted over ranked constraints.

*Keywords:* Optimality theory; Harmonic grammar; Phonology; Generative linguistics; Connectionism; Statistical learning; Language typology; Language learning

---

## 1. Introduction

Generative linguistics (GL) aims to formally characterize the set of possible human languages (Chomsky, 1957). 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 midpoint. A successful generative theory 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 the last pattern is expressible with simple mathematics, and many real stress patterns may seem much more complex. The formal structure of linguistic

---

Correspondence should be sent to Joe Pater, Department of Linguistics, University of Massachusetts, Amherst, MA 01003. E-mail: pater@linguist.umass.edu

connectionism (LC) is based on mathematical analogs of neural activity (e.g., Rumelhart & McClelland, 1986; Smolensky & Legendre, 2006). The aims of LC are diverse, but there is a central focus on the modeling of learning and other cognitive processes. GL and LC are often thought incompatible, as some versions of LC can learn and represent implausible linguistic systems (Pinker & Prince, 1988; see Drescher, 1994 for stress examples in a critique of Gupta & Touretzky, 1994, and see also Prince, 1993; Prince, 2007a on Goldsmith, 1994). From the perspective of GL, such theories are insufficiently *restrictive* in their predictions about the typology of natural languages.

Within GL, the typological fit of a theory of language is taken as a basic criterion of its success. From other perspectives, one might question this criterion, and also question the relevance of linguistic typology to cognitive science. It is not obvious that an adequate account of how an individual learns and uses language must take into account the distinction between attested and implausible linguistic systems: What real cost is there if the theory allows an individual to learn or use language types that a linguist might deem implausible? However, if one has as a goal a theory of learning and/or processing that is flexible enough to accommodate the full range of attested languages, then carefully formalized accounts of the systems of those languages should be an important source of information in constructing that theory. Such formal linguistic description is a central concern of GL. Furthermore, theories of language learning and processing are being used increasingly in the explanation of typological generalizations (for a range of viewpoints, see Hawkins, 2004 and Newmeyer, 2005 on syntax; and Bybee, 2001 and Moreton, 2008 on phonology). Therefore, the study of linguistic typology and the study of linguistic cognition are strongly mutually informing domains of research.

By combining aspects of earlier approaches to GL with aspects of LC, Optimality Theory (OT; Prince & Smolensky, 1993/2004, 1997) provides a framework for linguistic analysis that has seen broad application, especially in phonology, the study of the sound systems of the world's languages (see Kager, 1999; McCarthy, 2002, 2004 for overviews of research in OT). It has also had a significant impact on the study of language learning, again, especially in phonology (see Barlow & Gierut, 1999; Boersma & Levelt, 2003; Kager, Pater, & Zonneveld, 2004). OT characterizes linguistic systems in terms of constraints that place restrictions on the forms of a language. The types of linguistic representation that these constraints evaluate, and in many cases the constraints themselves, come from earlier work in GL. However, OT differs from most versions of GL 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.

As a version of GL with constraint interaction, OT creates an important bridge to other branches of cognitive science (Prince & Smolensky, 1993/2004, 1997; Smolensky, 2006a; Smolensky & Legendre, 2006). From the viewpoint of connectionist and statistical approaches to cognition, the general notion that one constraint can be overruled by another is a natural assumption, much more so than the contrasting assumption that an active constraint must be fully satisfied within its domain of application (as is traditional in GL; see further section 3.1). A specific connection is that an OT model of grammar can be

straightforwardly transformed into a grammatical model that is compatible with learning algorithms that have been developed for neural modeling and statistical approaches to cognition; see section 4.1 for discussion.

This transformation of OT grammars involves the replacement of OT's ranked constraints with numerically weighted ones. Thus, an even stronger bridge between GL and other research in cognitive science might be built by using weighted constraint grammars for the generative analysis of linguistic systems, as in OT's predecessor, Harmonic Grammar (HG; Legendre, Miyata, & Smolensky, 1990; Smolensky & Legendre, 2006; see also Goldsmith, 1990, 1991, 1993a, and other papers collected in Goldsmith, 1993b). According to Prince and Smolensky (1993/2004, 1997), Legendre, Sorace, and Smolensky (2006) and Smolensky (2006a), the stumbling block for this approach would be restrictiveness: A version of OT with weighted constraints would not properly distinguish between possible and impossible linguistic systems. Prince and Smolensky (1993/2004, 1997) cite a hypothetical interlocutor as having the following "fear of optimization":

- (1) *Loss-of-restrictiveness*: "In order to handle optimality, you must use numbers and use counting. . . The result will be a system of complicated trade-offs. . . giving tremendous descriptive flexibility and no hope of principled explanation. Therefore, the main goal of generative grammatical investigation is irredeemably undermined."

Their response (p. 233) is that this "[c]oncern is well-founded," but that in OT "recourse to the full-blown power of numerical optimization is not required...." In this journal, Smolensky (2006a) states that "[c]onstraint interaction in human grammars is more restricted than that permitted by arbitrary numerical constraint strengths. The more restricted theory of constraint interaction that yields empirically valid typologies is [OT]...."

These claims about the relative restrictiveness of HG and OT are accompanied by little explicit comparison of differences between the typological predictions of weighted and ranked constraints. Prince and Smolensky (1993/2004) present no examples of unattested linguistic patterns generated just by constraint weighting, and in fact, the only such case in the published literature appears to be one in Legendre et al. (2006), discussed in section 3.3 below.

The main point of the present article is that the predictions of weighted constraints for linguistic typology are often more restrictive than they may at first appear. The discussion is structured around the nature of *trade-offs* between constraints, which as Prince and Smolensky imply in the passage cited in (1), are key to understanding the similarities and differences between HG and OT (see further Prince, 2002; Legendre et al., 2006; Tesar, 2007). In the simplest kind of trade-off, the satisfaction of one of two constraints leads to the violation of the other. When constraint violations trade off in this symmetric fashion, assigning either greater positive weight or higher rank to one of the constraints leads to the same result. Section 2 starts by showing how a small set of phonological constraints produces identical typological predictions in OT and in a version of HG with positively weighted constraints. It then goes on to elaborate on the

equivalence between the two theories when trade-offs are symmetric, using an example that Prince and Smolensky (1997) in fact present as an illustration of the special properties of OT's ranked constraints.

Differences between HG and OT can emerge when constraint violations trade off asymmetrically; the simplest case is when two constraint violations can be avoided by violating only one conflicting constraint. An asymmetric trade-off provides a necessary condition for a gang effect, or cumulative constraint interaction, in which a constraint is satisfied at the cost of  $n$  violations of some lower valued constraint(s), but not  $n + 1$ . Prince and Smolensky introduce the term *strict domination* to refer to the inability of ranked constraints to express cumulative interaction. Section 3 illustrates cumulative constraint interaction with an example from the phonology of Japanese. On the issue of the relative restrictiveness of HG and OT, this section makes three points. The first is that it would be a mistake to refer to OT as inherently more restrictive than HG. Because weighted constraints can produce linguistic patterns that escape rankings of the same constraints, HG may capture attested languages using a constraint set that differs from an OT one. With different constraint sets in HG and OT theories, there is no necessary relationship of relative restrictiveness. The second point is that an asymmetric trade-off is not a sufficient condition for an HG–OT difference; this discussion is based in part on an asymmetric trade-off that Prince and Smolensky (1993/2004) take as evidence for strict domination. The third point is that the extent to which HG does indeed produce unattested linguistic patterns depends both on the types of constraints that are adopted and on the ways in which candidate representations are generated and evaluated.

Section 4 uses the Japanese example to briefly illustrate and extend existing arguments for the replacement of OT's ranked constraints with weighted ones: that the resulting model of grammar can be adapted relatively straightforwardly to deal with various types of noncategorical linguistic phenomena, and that it is compatible with well-understood algorithms for learning and other computations. As GL and LC have the shared goal of explaining how humans come to acquire their knowledge of language, the strengths of HG in this area are of considerable importance. This section also briefly addresses the issue of the origin of the constraints. Adapting a proposal from OT's generative predecessors (Chomsky's 1981 Principles and Parameters theory), Prince and Smolensky claim that the OT constraint set is universal, and that languages vary only in how the constraints are ranked. The postulation of universal constraint set is often interpreted as a claim that these constraints are innate (although cf. Smolensky, 2006a: fn. 8). Section 4.2 reviews a body of work in OT and HG that assumes that the constraints are in fact learned. It outlines a program of research in which universality is maintained as a useful heuristic for typological study in GL, with the adoption of weighted constraints aiding the interdisciplinary development of theories of learning and other types of linguistic cognitive processing. Ultimately, this interdisciplinary work holds the promise of creating fully formalized theories of linguistic typology that posit an innate component quite different from that of a universal constraint set.

## 2. Optimization, trade-offs, and restrictiveness

### 2.1. A simple phonological typology in OT and HG

We start with a simple example of how differences between linguistic systems are analyzed in OT and HG, to introduce linguists to the weighted constraint formalism, and to introduce nonlinguists to the analytic techniques and terminology of GL, and especially of OT. The example concerns the voicing of final obstruent consonants. In general, the obstruents are just those consonants that can be distinguished by whether they are voiced or not. Voiced obstruents are produced with vibration of the vocal cords (e.g., [b], [d], [g], [z], [v]; square brackets enclose phonetic symbols), in contrast to voiceless ones (e.g., [p], [t], [k], [s], [f]). In some languages, obstruents at the end of words are strictly voiceless: Obstruent voicing is banned in this position, even though it is permitted in other parts of the word. Languages with such final devoicing include Russian, German, and Thai, while languages with final voiced consonants include English, French, and Spanish.

This cross-linguistic variation can be analyzed in terms of the interaction of two basic types of OT constraint. The first is an *output* constraint that states the restriction against final voiced obstruents (output constraints are often termed “markedness constraints”; the distinction between input and output will be introduced shortly). The restriction against word-final voiced obstruents usually holds more broadly, including all consonants that occur in the final portion of a syllable. The portion of the syllable that follows the vowel is called the *coda*, and the constraint can thus be stated as a restriction against voiced obstruents in syllable codas (the asterisk in the constraint name indicates that the structure is penalized).

(2) \*CODA-VOICE

Every obstruent in coda position is voiceless

In languages like Russian, this constraint is satisfied,<sup>1</sup> and in languages like English, it is violated.

Under the view that constraints are universal, a language in which \*CODA-VOICE is violated is one in which a conflicting constraint has a higher position in the language-specific ranking. Output constraints often conflict with *faithfulness* constraints, which enforce similarity between levels of representation. In phonology, the basic levels of representation are the observed *surface* representations ( $\approx$ pronounced forms), and hypothesized *underlying* representations ( $\approx$ stored lexical representations). Underlying representations that diverge from surface forms are often hypothesized by the linguist, and presumably by the language learner, on the basis of alternations in the form of morphemes across phonological contexts. For example, in Russian, the genitive form of nouns (which gives a possessive meaning) is constructed by adding the suffix /-a/. The underlying voicing specification of the final consonant of the root morpheme can be observed in the genitive form, as it is protected from devoicing.<sup>2</sup> As shown in (3), the underlying representations are placed in angled brackets, to contrast with the surface forms in square brackets.

(3) *Russian final devoicing*

Nominative	Genitive	English gloss
/glaz/	/glaz-a/	'eye'
[glas]	[glaza]	

In OT, faithfulness constraints apply between an *input* and candidate *output* representations, while, as their name implies, output constraints assess only output forms. In phonology, the input is an underlying representation, and the output is a surface representation. The faithfulness constraint at issue demands that a consonant be consistently voiced or voiceless in its input and output forms (McCarthy & Prince, 1999).

## (4) IDENT-VOICE

Every consonant's voicing specification is identical in input and output

In a language like Russian, \*CODA-VOICE is ranked above IDENT-VOICE (written \*CODA-VOICE  $\gg$  IDENT-VOICE), and in a language like English, the ranking is IDENT-VOICE  $\gg$  \*CODA-VOICE.

The evaluation of the candidate set by the constraints can be illustrated in a *tableau* like the one in (5). The input is shown in the upper left cell, and output candidates are listed in the cells beneath. Prince and Smolensky propose that a universal *Gen* function takes an input and provides a set of output candidates. The language-specific constraint ranking then selects the optimal (input, output) pair from the candidate set. For our example, *Gen* must provide output candidates with either a voiced or voiceless obstruent for each of the input's voiced obstruents. The candidates in (5) show all the logically possible combinations in this regard for the input /bad/. This hypothetical form stands in for any word that has the relevant phonological properties (i.e., voiced obstruents in a coda and non-coda position). Constraints are listed in columns in rank order. Under each constraint are asterisks indicating the violations incurred by each candidate. \*CODA-VOICE assigns a violation mark for every coda voiced obstruent. Both [bad] and [pad] have a voiced coda [d], and hence violate the constraint, while the other candidates have voiceless [t] and avoid violations. IDENT-VOICE assigns a mark for every obstruent whose voicing is changed between input and output.

(5) *Final devoicing in OT*

/bad/	*CODA-VOICE	IDENT-VOICE
bad	*	
pad	*	*
 [bat]		*
pat		**

Because \*CODA-VOICE is ranked above IDENT-VOICE, the voiceless coda [t] in [bat] is preferred over the faithful coda [d] of [bad]. Evaluation by ranked constraints is usually defined procedurally.<sup>3</sup> The first step takes the highest ranked constraint, finds the candidate(s) with the fewest violation marks, and eliminates all others from consideration. In (5), this eliminates the first two candidates due to their violations of \*CODA-VOICE. This filtering procedure iterates down the hierarchy, until there is only one candidate left, which is declared the optimum. In (5), [bat] becomes optimal after IDENT-VOICE eliminates the output [pad]. Optimality is indicated with the pointing finger. Here, the optimal output is the surface phonological representation, so it is enclosed in square brackets.

When we reverse the ranking, [bad] becomes optimal. This is illustrated in (6), in which all output candidates but [bad] are eliminated by IDENT-VOICE.

(6) *Final voiced obstruents permitted in OT*

/bad/	IDENT-VOICE	*CODA-VOICE
 [bad]		*
pad	*	*
bat	*	
pat	**	

In OT, the predicted typology of possible languages corresponds to the sets of optima selected by some ranking of the constraint set. Given that we have exhausted the possible rankings of this small constraint set, the resulting prediction is that for an input like /bad/, no language will ever have either [pad] or [pat] as an output. The second of these is in fact the correct output for a language with no voiced obstruents at all. Such context-free devoicing becomes a possible result if the constraint set includes a general \*VOICE constraint ('Obstruents are voiceless'), as the tableau in (7) shows (\*CODA-VOICE is omitted because its ranking is irrelevant to the outcome).

(7) *Context-free devoicing in OT*

/bad/	*VOICE	IDENT-VOICE
bad	**	
pad	*	*
bat	*	*
 [pat]		**

In the languages that contain voiced obstruents, the ranking would be reversed (IDENT-VOICE ≫ \*VOICE).

While there are languages with final devoicing, and languages that lack voiced obstruents entirely, there are no attested languages in which voicing is banned only in syllable-initial position. This gap in the typology is captured by our expanded constraint set, as there is still

no ranking that can make [pad] the optimal output for /bad/. Its consistent suboptimality is due to the fact that none of the constraints prefer it over the final devoicing output [bat]. Both of these violate IDENT-VOICE once, and [pad] violates in addition \*CODA-VOICE. This relationship in violation profiles, where one member of a candidate set has a proper superset of the violations of another, is referred to as harmonic bounding (Samek-Lodovici & Prince, 1999): (/bad/, [bat]) harmonically bounds (/bad/, [pad]).

Thus, of the four output candidates for /bad/, three can be made optimal in OT: [bad], [bat], and [pat], but not [pad]. We will now see that the same typology emerges from a version of OT with weighted constraints. In HG, a candidate evaluated in terms of its *harmony* ( $H$ ), which is calculated by a simple linear equation (see Smolensky & Legendre, 2006 on the intellectual history of HG harmony). Each constraint  $C_k$  ( $k = 1, \dots, K$ ) is associated with a real numbered weight  $w_k$ . The candidate's violation or satisfaction score on each constraint  $s_k$  is multiplied by the weight, and the results are then summed:

(8) *Harmony*

$$H = \sum_{k=1}^K s_k w_k$$

Prince and Smolensky (1993/2004: 236) point out that OT evaluation could use such weighted sums to determine optimality (see also Goldsmith, 1991: fn. 10). As in Smolensky and Legendre (2006), I assume that the optimum is a candidate with maximal harmony in its candidate set, with a candidate's harmony defined as in (8). In most of this paper, I will impose the further restriction that the optimum must have greater harmony than any of its competitors, thus ruling out ties for optimality (if candidates have different violation profiles). This brings the theory closer to the original version of OT, in which there is usually only a single optimum for each candidate set in a language. The imposition of a uniqueness condition on optima is an idealization that facilitates typological study and allows us to isolate the effects of weighting and ranking; see section 4.1 on versions of HG and OT that capture variation between optima within a candidate set in a single language. Following Legendre et al. (2006), I convert OT violation marks into negative integers, which indicates that they assign penalties; this allows a possible extension to constraints that assign positive numbers for satisfaction. For reasons to be discussed shortly, I limit weights to nonnegative values. When constraints assign negative scores, and weights are nonnegative, the optimum has the value closest to zero, that is, the lowest penalty.

The HG tableau in (9) adds a row to show the weights of each constraint, and a column to show the harmony scores of each candidate. Because \*CODA-VOICE has greater weight than IDENT-VOICE, a candidate that incurs one violation of that constraint, like [bad], receives a greater weighted penalty ( $2 \cdot (-1) = -2$ ) than a candidate that has a single violation of IDENT-VOICE, like [bat] ( $1 \cdot (-1) = -1$ ). As (/bad/, [bat]) has the highest harmony in the candidate set, it is optimal.

(9) *Final devoicing in HG*

	2	1	
/bad/	*CODA-VOICE	IDENT-VOICE	
bad	-1		-2
pad	-1	-1	-3
☞ [bat]		-1	-1
pat		-2	-2

Such coda devoicing will result with any set of positive weights for these constraints that respects the strict inequality in (10) where  $w(C)$  is the weight of constraint C:

(10) *Weighting condition for final devoicing language*

$$w(*\text{CODA-VOICE}) > w(\text{IDENT-VOICE})$$

Conversely, voiced obstruents are allowed in coda position (e.g., the pair (/bad/, [bad]) is optimal) if the direction of the strict inequality is reversed.

(11) *Weighting condition for a language that allows final voiced obstruents*

$$w(\text{IDENT-VOICE}) > w(*\text{CODA-VOICE})$$

Parallel to the OT result, there is no positive set of weights for these constraints that will make either [pad] or [pat] beat all of their competitors in the tableau in (9), as the two weighting conditions result only in either [bat] or [bad] being optimal. Again as in OT, by adding \*VOICE to the HG constraint set, we can make [pat] optimal. It will win under the following weighting condition, which yields the same result as the corresponding ranking in (7).<sup>4</sup>

(12) *Weighting condition for a language with context-free devoicing*

$$w(*\text{VOICE}) > w(\text{IDENT-VOICE})$$

More generally, we know that any candidate that can be made optimal by some ranking of a set of constraints can also be made optimal by some weighting of that same set of constraints. Any finite set of forms that are optimal in their candidate sets under some OT ranking can be optimal with some HG weighting of the same constraints (Legendre et al., 2006; Prince, 2002; Prince & Smolensky, 1993/2004: 236). Therefore, we knew from the outset that the three outputs for /bad/ that are possible OT optima, [bad], [bat], and [pat], could be made optimal by using our OT constraint set in HG. But what about the initial devoicing outcome (/bad/, [pad]), which was ruled out in OT?

In HG with strictly positive weights, [pad] cannot be the optimal output for /bad/ for the same reason as in OT: because the final devoicing candidate [bat] is preferred by

\*CODA-VOICE, and there is no constraint that prefers [pad]. More generally, Prince (2002) shows that in a version of HG that limits weights to positive values, harmonic bounding relationships involving proper subsets of violation marks are preserved. If negative weights were allowed, however, [pad] could be made to win. For example, if the signs of the weights in the tableau in (9) were reversed from (2,1) to (-2,-1), then the signs of the harmony scores would be reversed too, and (/bad/, [pad]) would have greater harmony than any of its competitors. This illustrates the general point that a constraint that can be weighted both negatively and positively can both punish and reward the same structure; if we want OT-like behavior from HG constraints, their weights must be kept on one side of zero.

Thus, in this simple case, the predicted typology produced by constraint ranking and positively weighted constraints is identical; the same subset of candidates can be made to beat their competitors in either theory.

One might well wonder how this HG approach to typology could scale up to less idealized cases, with more inputs, output candidates, and constraints. Under the standard assumption that there is a finite constraint set, there is a finite space of possible rankings, so one can, in principle, examine the predictions of all those rankings. How can one examine the unbounded space of possible weightings in HG? The answer relates to the fact that even examining the space of possible rankings becomes impractical when constraint sets get large enough (and this happens very quickly, given that number of rankings is the factorial of the number of constraints). Instead, it is usually more efficient to examine the sets of possible optima across candidate sets. In OT, whether optima are consistent, that is, whether they can be made optimal by a single ranking, can be determined using the Recursive Constraint Demotion Algorithm (Tesar & Smolensky, 1998, 2000). This algorithm forms the basis for the calculation of typologies in the software package OT-Soft (Hayes, Tesar, & Zuraw, 2003). In HG, Potts et al. (2009) show that consistency between optima can be determined using the simplex algorithm of linear programming; this forms the basis for the HG typology calculations in OT-Help (Becker, Pater, & Potts, 2007). This application takes weighting conditions of the type I have been discussing here and uses the simplex algorithm to determine whether they are consistent across candidate sets (see section 3.1 for examples of multiple candidate sets in a language). If the conditions are consistent, a set of weights is found that produces an HG analysis of the language. This application greatly aids linguistic analysis and study of typology in HG, and it forms an example of how HG's linear model of grammar allows the use of existing, well-understood algorithms for computational implementations (see further section 4.1).

## 2.2. *Symmetric trade-offs and absence of gang effects*

We now turn to a further case in which the typological predictions of OT and HG with this constraint set continue to converge, even though it might seem at first that they could differ. HG allows for gang effects, in which the number of violations of one or more constraints with lower weight determines whether a constraint with higher weight is satisfied. A gang effect is illustrated in (13). Because the summed weights of CONSTRAINT-2 and CONSTRAINT-3 are greater than that of CONSTRAINT-1, the candidate

that violates only CONSTRAINT-1 is chosen as optimal. If the violation profiles were changed such that Out-12 only violated one of the lower weighted constraints, it would be optimal under these weights.

(13) *An abstract gang effect in HG*

		1.5	1	1	
	In-1	CONSTRAINT-1	CONSTRAINT-2	CONSTRAINT-3	
☞	Out-11	-1			-1.5
	Out-12		-1	-1	-2

With the OT ranking  $\text{CONSTRAINT-1} \gg \text{CONSTRAINT-2}, \text{CONSTRAINT-3}$ , Out-12 would be the optimal output, no matter how many violations of CONSTRAINT-2 and CONSTRAINT-3 it incurred.

Given the possibility of gang effects, one might imagine that HG could produce a pattern in which one voiced coda obstruent is tolerated, but words with a greater number of codas have devoicing. This is in fact impossible using the constraints \*CODA-VOICE and IDENT-VOICE because of the way their violations trade off (see Prince, 2002 on one-to-one trade-offs in HG-OT translations). Each potential voiced coda incurs either one violation of \*CODA-VOICE if it is voiced in the output, or one violation of IDENT-VOICE if it is voiceless. Thus, for every violation of \*CODA-VOICE that a candidate avoids, it will incur one violation of IDENT-VOICE, and vice versa, as illustrated in (14). In other words, violations trade off *symmetrically*, in contrast to the *asymmetric* trade-off in (13). Periods in the output candidates indicate syllable boundaries; word-internal codas precede the periods.

(14) *Symmetric trade-off*

/dagbad/	*CODA-VOICE	IDENT-VOICE
dag.bad	-2	
dag.bat	-1	-1
dak.bat		-2

If  $w(\text{*CODA-VOICE}) > w(\text{IDENT-VOICE})$ , all coda obstruents will be voiceless (e.g., [dak.bat]) and if  $w(\text{IDENT-VOICE}) > w(\text{*CODA-VOICE})$ , all underlying voiced obstruents will surface as voiced (e.g., [dag.bad]). The candidate [dag.bat] is *collectively* harmonically bounded (Samek-Lodovici & Prince, 1999) by the competitors shown in this tableau: Under no weighting will it beat them both.<sup>5</sup> With this constraint set and this theory of constraint interaction, there is no way for the number of voiced obstruents in a word to influence whether any one of them devoices.

This result would not hold if we set a numerical cut-off, or threshold, on well-formedness. For example, if we defined well-formed (input, output) pairs as those with Harmony greater than  $-1.5$ , we could produce the one voiced coda maximum with \*CODA-VOICE weighted at 1, as shown in (15). Well-formed pairs are indicated with a checkmark, and the ill-formed one with an asterisk.

(15) *Evaluation of (input, output) pairs in a threshold system*

	1	<i>H</i>
	*CODA-VOICE	
✓ (/bad/, [bad])	-1	-1
✓ (/dagbad/, [dag.bat])	-1	-1
* (/dagbad/, [dag.bad])	-2	-2

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

In making the case for OT to an interdisciplinary audience, Prince and Smolensky (1997: 1604) draw the generalization that:

- (16) 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.

They attribute this type of phenomenon to strict domination property of ranked constraints. As an example, they discuss the interaction of NoCODA and PARSE: NoCODA is an output constraint against codas, and PARSE is a faithfulness constraint that demands that input segments be parsed into output syllable structure; a consonant that is unparsed is unpronounced ( $\approx$ deleted).<sup>6</sup> Prince and Smolensky (1997: 1606) state that:

- (17) Domination is clearly “strict” in these examples: No matter how many consonant clusters appear in an input, and no matter how many consonants appear in any cluster, [the grammar with NoCODA $\gg$ 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 $\gg$ 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 this passage, Prince and Smolensky offer as an illustration of strict domination the observation that the number of consonant clusters (i.e., potential codas) does not affect

whether deletion occurs or not. The table in (18) shows that the trade-offs across candidates' violation profiles have the same symmetry as the coda devoicing case in (14).<sup>7</sup> Unparsed segments are placed between angled brackets; as they are unparsed, they do not violate NoCODA.

(18) *Symmetric trade-off*

/dagbadga/	NoCODA	PARSE
dag.bad.ga	-2	
da<g>.bad.ga	-1	-1
da<g>.ba<d>.ga		-2

Regardless of the number of potential codas, there are only two outcomes: If  $w(\text{PARSE}) > w(\text{NoCODA})$  the language will permit codas (as in [dag.bad.ga]), and if  $w(\text{NoCODA}) > w(\text{PARSE})$ , all potential codas will fail to be pronounced (as in [da<d>.ba<d>.ga]). Thus, strict domination is irrelevant to the irrelevance of the number of potential codas; the lack of that number's effect on whether deletion applies is a prediction of an optimization system with these constraints, be they ranked or weighted.<sup>8</sup>

3. Asymmetric trade-offs

3.1. *Nonvacuous gang effects*

As an example of an asymmetric trade-off that leads to an HG–OT difference, that is, a *nonvacuous* gang effect, we can consider a slightly more complicated example in the phonology of obstruent voicing. In Japanese, only a single voiced obstruent ([b], [d], [g], [z]) is usually permitted in a word (see Ito & Mester, 1986, 2003 for analyses in GL). This restriction is usually termed Lyman's Law (Lyman, 1894). In loanwords, however, multiple voiced obstruents are permitted (Kawahara, 2006; Nishimura, 2003, 2006; all data are from the former source):

(19) *Violations of Lyman's Law in loanwords*

[bagi:] 'buggy'	[bagu] 'bug'
[bogi:] 'bogey'	[dagu] 'Doug'
[bobu] 'Bob'	[giga] 'giga'

Japanese also has a restriction against obstruent voicing in geminates (consonants with long duration, here marked with a colon). But again, in loanwords, voiced geminate obstruents are permitted:

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

[web:u] ‘web’	[wip:u] ‘whipped (cream)’
[sunob:u] ‘snob’	[sutop:u] ‘stop’
[hab:uru] ‘Hubble’	[kap:uru] ‘couple’
[kid:o] ‘kid’	[kit:o] ‘kit’
[red:o] ‘red’	[autoret:o] ‘outlet’
[hed:o] ‘head’	[met:o] ‘helmet’

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

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

[gud:o] ~ [gut:o] ‘good’	[dog:u] ~ [dok:u] ‘dog’
[bed:o] ~ [bet:o] ‘bed’	[bag:u] ~ [bak:u] ‘bag’
[dored:o] ~ [doret:o] ‘dredlocks’	[bud:a] ~ [but:a] ‘Buddha’
[bad:o] ~ [bat:o] ‘bad’	[dorag:u] ~ [dorak:u] ‘drug’
[deibid:o] ~ [deibit:o] ‘David’	[big:u] ~ [bik:u] ‘big’

According to Nishimura (2003, 2006) and Kawahara (2006), such devoicing is judged unacceptable in the word types illustrated in both (19) and (20).

In HG, this devoicing pattern can be analyzed as being due to two independently motivated constraints. This analysis draws on Nishimura’s (2003, 2006) account using Smolensky’s (2006b) OT with Local Conjunction; the possibility of an HG reanalysis was suggested by Shigeto Kawahara (p.c.). The first constraint expresses a cross-linguistically common ban against voiced obstruent geminates, which can be held responsible for their absence in native Japanese words. The constraint \*V<sub>CE</sub>-GEM is violated by each one of these in the output string. The other constraint is OCP-VOICE, which Ito and Mester (1986) propose to account for the Lyman’s Law restriction in native Japanese words. This constraint penalizes every sequence of voiced obstruents, even ones separated by any number of other segments. For example, [dored:o] has one violation of OCP-VOICE ([d...d:]), and [deibid:o] has two ([d...b], [b...d:]), while [doret:o] has no violations of OCP-VOICE, and [deibit:o] has one. In (22), the weight of IDENT-VOICE is greater than that of each of OCP-VOICE and \*VOICE-OBS, so that a pair of voiced obstruents is permitted (see the first tableau), as is a voiced geminate (see the second tableau). The last tableau contains the asymmetric trade-off necessary for the HG-OT difference: A single violation of IDENT-VOICE trades off against violations of \*V<sub>CE</sub>-GEM and OCP-VOICE. As the summed weight of OCP-VOICE and \*V<sub>CE</sub>-GEM is greater than that of IDENT-VOICE, the geminate devoices in the presence of another voiced obstruent.

(22) Japanese loanword devoicing as cumulative constraint interaction

	1.5	1	
/bobu/	IDENT-VOICE	OCP-VOICE	
☞ [bobu]		-1	-1
bopu	-1		-1.5

  

	1.5	1	
/web:u/	IDENT-VOICE	*VCE-GEM	
☞ [web:u]		-1	-1
wep:u	-1		-1.5

  

	1.5	1	1	
/dog:u/	IDENT-VOICE	*VCE-GEM	OCP-VOICE	
dog:u		-1	-1	-2
☞ [dok:u]	-1			-1.5

In this gang effect, or *cumulative constraint interaction*, we have two constraints ganging up to overcome a third one with higher weight. Due to strict domination, OT cannot express this sort of constraint interaction. In this case, if IDENT-VOICE were ranked above each of \*VCE-GEM and OCP-VOICE, as required for the first two tableaux, the optimal output for /dog:u/ would be [dog:u], rather than [dok:u].

Although many other linguistic patterns could be analyzed in terms of cumulative constraint interaction (in phonology, see e.g., Guy, 1997 as a precedent), this example provides particularly striking evidence for HG because OCP-VOICE and \*VCE-GEM are independently motivated in the phonology of Japanese. Section 4 shows that this pattern emerges as a stage of acquisition in an HG model of language learning. Here I discuss how this example distinguishes HG-weighted constraints from the traditional generative view of constraints as inviolable, and from the OT view of them as ranked. As noted above, the OCP-VOICE constraint was posited by Ito and Mester (1986) to account for the fact that the native vocabulary is restricted by Lyman’s Law, and for the morphological alternations that show its effects. In GL outside of OT and HG, an active constraint is a true statement about the domain in which it applies; in other words, an active constraint is inviolable. Given the examples of loanwords with pairs of voiced obstruents like [bobu], OCP-VOICE would have to be considered inactive, at least for loanwords (see Ito & Mester, 1995). The prediction, then, is that OCP-VOICE should not have any effect on the devoicing of geminates. Similarly, if OCP-VOICE is outranked by IDENT-VOICE for the OT account of [bobu]-type words, then as we have seen, it cannot participate in the devoicing of geminates. Nishimura’s (2003, 2006) discovery of the cumulative effect of OCP-VOICE and \*VCE-GEM in Japanese loanwords thus falsifies these predictions of traditional GL and of OT, and weighs in favor of HG.

The predictions of these frameworks depend, of course, on the contents of the constraint sets. With a different constraint set, one could analyze the Japanese case with either inviolable constraints or ranked ones. Nishimura (2003, 2006) posits a conjoined version of OCP-VOICE and \*VCE-GEM, which is violated *iff* a word contains violations of both constraints. Local conjunction of this type can probably account for many, if not all, observed patterns of cumulative constraint interaction. There are two issues. First, there is no learning theory for local conjunction (Smolensky, 2006b), and hence no explanation for why learners should pass through a local conjunction stage (cf. section 4). Second, as Pater (2009a) and Potts et al. (2009) show, local conjunction yields unattested systems that HG does not reproduce. The exploration of the relationship between HG and OT with constraint conjunction is an important topic for further research (see also Legendre et al., 2006), but it goes beyond the scope of the present paper's comparison of the original version of OT with HG. As the goal here is to establish a general picture of the relationship between the frameworks, comparing the present analysis with Kawahara's (2006) OT account, which employs different nonconjoined constraints, would also take us too far afield.

The comparison of HG and OT with the present constraint set establishes the point that HG can generate attested patterns that fall out of reach when the same constraints are ranked. This allows us to draw two important conclusions. First, weighted constraint interaction can in principle permit a smaller set of more general constraints than that used in OT analyses of the same data. Much of the appeal of ranked versus inviolable constraints in fact comes from a similar ability of ranking to reduce phenomena to more general principles (Prince and Smolensky, 1993/2004: secs. 4, 10.3). Second, because HG can account for attested patterns with a constraint set that differs from the one needed in OT, it is likely that any fleshed out theory of some set of linguistic phenomena in HG will be in some ways less restrictive, and in some ways *more* restrictive, than the comparable OT one. For further discussion, see Potts et al. (2009).

### 3.2. *Vacuous gang effects*

In order to provide a reasonably accurate general depiction of the relationship between the patterns produced by weighted and ranked constraint interaction, it is important to emphasize that while an asymmetric trade-off is a necessary condition for an HG–OT difference, it is not a sufficient one. Consider, for example, the asymmetric trade-off involving some constraints and candidates from section 2.1 (here and in the following tableau, I omit the shared violation of \*VOICE incurred by the [b]):

#### (23) *Asymmetric trade-off with voicing constraints*

/bad/	IDENT-VOICE	*VOICE	*CODA-VOICE
bad		-1	-1
bat	-1		

The reason that this violation profile does not lead to an HG–OT difference is that a gang effect between \*VOICE and \*CODA-VOICE would be vacuous. To see this, let us first apply a set of weights fitting the by-now familiar schema needed for the cumulative constraint interaction.

(24) *Coda devoicing as a gang effect*

	1.5	1	1	
/bad/	IDENT-VOICE	*VOICE	*CODA-VOICE	
bad		-1	-1	-2
 [bat]	-1			-1.5

Next, we need to seek the cases in which a violation of IDENT-VOICE can be traded off against a single violation of \*VOICE, and against a single violation of \*CODA-VOICE, in a way parallel to the Japanese example in (22). The problem is that the second of these does not exist, as every violation of \*CODA-VOICE entails a violation of \*VOICE, as in (24). Thus, we have only the scenario in which IDENT-VOICE trades off against a single violation of \*VOICE. As the single \*VOICE violation is insufficient to override IDENT-VOICE, devoicing is blocked syllable initially.

(25) *No initial devoicing*

	1.5	1	1	
/ba/	IDENT-VOICE	*VOICE	*CODA-VOICE	
 [ba]		-1		-1
pa	-1			-1.5

This is the final devoicing pattern we analyzed with \*CODA-VOICE having greater weight than \*VOICE, and with the OT ranking \*CODA-VOICE ≫ \*VOICE. The gang effect between \*CODA-VOICE and \*VOICE is vacuous, as it fails to produce a new set of optima, that is, a new language.

Many constraints in the OT literature stand in the specific-to-general relationship exemplified by \*VOICE and \*CODA-VOICE, in which the specific constraint assigns violations to a proper subset of the forms violated by the general constraint (see, e.g., de Lacy, 2006). The gang effect between any of these pairs of constraints is always vacuous, as it is indistinguishable from the pattern produced by obedience to the specific constraint alone. In all of these cases, then, asymmetric trade-offs will fail to yield HG–OT differences in typology (although see Jesney & Tessier, 2007 on advantages of HG for modeling language learning when faithfulness constraints are in a specific-to-general relationship).

As another type of example of a vacuous gang effect, we turn to a case of constraint interaction that Prince and Smolensky (1993/2004) point to as illustrating the difference between ranking and weighting. It involves the pair of candidates in their tableau (183A), which forms part of their analysis of Lardil final vowel truncation (Hale, 1973):

(26) *Tableau 183A*

	/yiliyili/	FREE-V	ALIGN	PARSE	NoCODA
☞	yi.li.yil.<i>		-1	-1	-1
	yi.li.yi.li.	-1			

The constraints PARSE and NoCODA were introduced in section 2.2: They are violated in the first candidate because the final vowel is unparsed and the last consonant is syllabified as a coda. These violations serve to satisfy the output constraint FREE-V, which demands that the word-final vowel be unparsed. Satisfaction of FREE-V also forces a violation of the constraint ALIGN, which requires the edge of the word to coincide with a syllable boundary. Prince and Smolensky make two comments about this tableau. The first is on p. 144:

- (27) 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 second is on p. 148:

- (28) *Strictness of strict domination.* In several examples the correct analysis violates many constraints, and its optimality rests crucially on the fact that competitors with a cleaner record overall happen to violate some single dominant constraint. Recall the discussion of /yiliyili/ in 7.3.2: a strong contender violating just one constraint is bested by an optimal parse violating three of the four less dominant constraints. This effect highlights the content of the central evaluative hypothesis, and sets the theory apart from others in which richer notions of ‘weighting’ and ‘trade-off’ are entertained.

It is in fact not clear how the Lardil example is meant to set ranking apart from weighting. First, we can obviously assign a set of weights to the constraints to pick the correct optimum, as shown in (29).

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

	4	1	1	1	
	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.

Less obviously, any gang effect between the three constraints violated by [yi.li.yil.⟨i⟩] would be vacuous. This is due to another kind of specific-to-general relationship that obtains between FREE-V and both ALIGN and PARSE: Any candidate that satisfies FREE-V necessarily violates ALIGN and PARSE, but not vice versa (Prince & Smolensky, 1993/2004: 7.2.1). Because FREE-V satisfaction entails the violation of these constraints, a gang effect that involves ALIGN and/or PARSE with NoCODA in blocking deletion would be vacuous. It would be vacuous because of the failure of these constraints to provide a one-to-one trade-off between NoCODA and FREE-V. In the absence of this one-to-one trade-off, there would be no occasion for the lower weight of NoCODA than FREE-V to show its effect. To put it differently, the sum of the effects of ALIGN and/or PARSE with NoCODA in forcing the violation of FREE-V would be the same as the effect of NoCODA alone. As this gang effect is vacuous, it does not produce a divergence in the typological predictions of HG and OT.

### 3.3. Unbounded trade-offs and locality

As mentioned in the introduction, Legendre et al. (2006) provide what seems to be the only published example of an unattested linguistic system produced by HG but not OT. The example involves what can be referred to as an unbounded trade-off: Satisfaction of one constraint can require a potentially unbounded number of violations of another. The constraints at issue are ones that determine stress placement. The first requires that a particular kind of syllable—a *heavy* one—be stressed (WEIGHT-TO-STRESS; Prince, 1990). Languages vary in which syllables fall into the heavy category; for the abstract example below, heavy syllables are ones that have codas. The other constraint penalizes a stressed syllable according to how far away it is from a word edge. This constraint was proposed by Prince and Smolensky (1993/2004) as one of a family of *Edgemo*st constraints, which were later reformalized as *Generalized Alignment* constraints by McCarthy and Prince (1993). The Alignment constraint ALIGN-HEAD-R demands that the main stress be at the right edge of the word, and it assigns a violation mark for each syllable that intervenes. For expository purposes, I adopt the name MAINSTRESSRIGHT from Legendre et al. (2006).

The tableau in (30), adapted from Legendre et al. (2006), compactly illustrates the unbounded trade-off. The syllable [ban] stands in for any heavy syllable, the coda-less syllable [ta] stands in for any nonheavy one, and the acute accent marks stress placement. The variable  $\sigma_n$  is a string of a number  $n$  of nonheavy syllables. As in Legendre et al. (2006), we only consider candidates with a single stress per word. Stress on the final syllable violates WEIGHT-TO-STRESS and satisfies MAINSTRESSRIGHT. Stress on the initial syllable satisfies WEIGHT-TO-STRESS and violates MAINSTRESSRIGHT once for every syllable separating it from the right edge of the word. Stress on any of the syllables in  $\sigma_n$  would be harmonically bounded by final stress, as they would add at least one violation of MAINSTRESSRIGHT without compensating improvement on WEIGHT-TO-STRESS. Therefore, we need only consider the two candidates in (30).

(30) *An unbounded trade-off*

/ban $\sigma_n$ ta/	WEIGHT-TO-STRESS	MAINSTRESSRIGHT
ban. $\sigma_n$ .tá	-1	
bán. $\sigma_n$ .ta		-1-n

As there is no theoretical upper bound on the size of words, there is theoretically no upper bound on the number of MAINSTRESSRIGHT violations that can be traded off against the single violation of WEIGHT-TO-STRESS.

Legendre et al. (2006) point out that in OT, there are only two possible languages, given by the two rankings. The number of syllables intervening between a nonfinal stressed heavy syllable and the edge of the word is irrelevant; either stress will fall on the rightmost syllable or on the heavy syllable.

(31) *Two languages in OT*

WEIGHT-TO-STRESS $\gg$ MAINSTRESSRIGHT	☞bán. $\sigma_n$ .ta
MAINSTRESSRIGHT $\gg$ WEIGHT-TO-STRESS	☞ban. $\sigma_n$ .tá

They also note that HG produces a theoretically infinite set of languages with these constraints. With appropriate weights, stress can be limited to a ‘‘window’’ of any number of syllables at the right edge of the word (see also Prince, 1993, 2007a on Goldsmith, 1994). In HG, the number of intervening syllables is crucial, as shown by the fact that this number is included in the weighting conditions in (32). Stress will fall on a nonfinal heavy syllable only if the weight of WEIGHT-TO-STRESS is greater than the weight of MAINSTRESSRIGHT times the number of syllables separating the heavy syllable from the right edge. For example, if  $w(\text{WEIGHT-TO-STRESS}) = 3.5$  and  $w(\text{MAINSTRESSRIGHT}) = 1$ , then a heavy syllable will get stressed if it is followed by three light syllables, but not four. No known language has such a four-syllable window.

(32) *An infinite typology in HG*

$w(\text{WEIGHT-TO-STRESS}) > (n+1) \cdot w(\text{MAINSTRESSRIGHT})$	☞bán. $\sigma_n$ .ta
$(n+1) \cdot w(\text{MAINSTRESSRIGHT}) > w(\text{WEIGHT-TO-STRESS})$	☞ban. $\sigma_n$ .tá

Taken on its own, this case is not particularly persuasive as an argument against HG as a framework for typological study. One problem is that there are, in fact, attested three-syllable windows (e.g., in Macedonian: Comrie, 1976; and Pirahã: Everett & Everett, 1984), and these cannot be generated by OT rankings of the set of constraints for stress in Prince and Smolensky (1993/2004) and McCarthy and Prince (1993) (see Hyde, 2007). One might take the ability of HG to account for them as a positive result and seek an explanation for the

absence of the larger windows, perhaps in terms of the size of words that a learner would need to hear to acquire the pattern (cf. Hammond, 1991), or in terms of the relative difficulty of acquiring the weight ratios needed to represent the pattern (cf. Prince, 1993: 91; Prince, 2007a: 41).<sup>9</sup>

Another problem is that gradient Alignment constraints are controversial, even in OT. In counting the distance between two portions of the representation, these constraints assign violation scores in an unusual manner, both in comparison with other OT constraints, and with ones elsewhere in GL (Bíró, 2003; Eisner, 1998; McCarthy, 2003; Potts & Pullum, 2002). For the other constraints discussed in this paper, which are typical of OT, each structure that violates the constraint gets just one violation mark. For a stressed syllable evaluated by a MAINSTRESSRIGHT, the number of violations depends on the distance from the edge of the word. Not only are the gradient Alignment constraints formally unusual, but they also produce undesired typological predictions, in OT as well as HG. Based on these considerations, McCarthy (2003) proposes a revised theory of OT constraints in which gradient Alignment constraints are banned. Removing gradient Alignment constraints from OT has the effect of also removing a large class of potential problem cases for a version of the theory with weighted constraints.

Unbounded trade-offs, and divergences between the typological predictions of OT and HG, can also emerge from the interaction of constraints that only assign a single violation per structure. In some situations, the satisfaction of an output constraint can require a number of faithfulness violations with no theoretical upper bound.<sup>10</sup> As an example, we can consider the interaction of NoCODA with the faithfulness constraint LINEARITY (McCarthy & Prince, 1999), which assigns a violation mark for every pairwise reordering of the segmental string.

(33) LINEARITY

If segment *x* precedes segment *y* in the input, *x* precedes *y* in the output

In (34), word-internal syllable boundaries are again indicated with periods, and NoCODA violations again occur when a syllable ends in a consonant (clusters are assumed to be split between syllables). Here we see that satisfaction of NoCODA can require two reorderings of the segmental string, as in the final candidate.

(34) *Asymmetric trade-off between NoCODA and LINEARITY*

/apekto/	NoCODA	LINEARITY
[a.pək.to]	-1	
[pa.ək.to]	-1	-1
[ap.ke.to]	-1	-1
[pa.ke.to]		-2

It is also possible to create strings in which only one violation of LINEARITY would be needed to satisfy NoCODA (e.g., /ekto/, [ke.to]), as well as ones in which any higher number is needed (e.g., /idapekto/ requires three violations, as in [di.pa.ke.to]). Appropriate weightings of the constraints can create systems in which NoCODA is satisfied at the cost of  $n$  violations of LINEARITY, but not  $n + 1$  violations, where  $n$  is any nonnegative integer. OT only produces two systems: one in which NoCODA is satisfied at any cost in terms of LINEARITY violations, and one in which even a single LINEARITY violation is worse than a violation of NoCODA.

McCarthy (2007) in fact provides this example as a case in which the standard version of OT produces the wrong result. Although languages do employ local pairwise reorderings of segments to satisfy output constraints like NoCODA (see Hume, 2001 for a survey), none use a double reordering of the type illustrated in the final candidate in (34), which would be optimal under the ranking NoCODA  $\gg$  LINEARITY (as well with weights respecting the condition  $w(\text{NoCODA}) > 2 \cdot w(\text{LINEARITY})$ ).

McCarthy shows that the correct typology is obtained in an alternative version of OT that Prince and Smolensky (1993/2004) call Harmonic Serialism. In this theory, Gen is limited to a single application of an operation; here it can produce a single pairwise reordering, but not two at once. Multiple applications of an operation can occur serially, if each one results in an improvement in harmony. The tableau in (35) shows the first candidate set that would be produced in a Harmonic Serialist derivation, which lacks the candidate with a double reordering. The faithful candidate harmonically bounds the others, so it would always be picked as optimal regardless of the constraint ranking. In Harmonic Serialism a derivation terminates when the faithful candidate is chosen; (35) is thus both the first and last step.

(35) *No double reordering in Harmonic Serialism*

	/apekto/	NoCODA	LINEARITY
☞	[a.pek.to]	-1	
	[pa.ek.to]	-1	-1
	[ap.ke.to]	-1	-1

McCarthy's solution extends to a weighted constraint version of the theory, as [a.pek.to] is equally guaranteed to win in (35) with any set of positive weights. It also eliminates the difference between OT and HG typology mentioned beneath (34), since in a serial version of either OT or HG, only single reorderings (e.g., /ekpo/, [ke.po]) can be used to satisfy NoCODA. Although a proper elaboration of a serial version of HG is impossible here, this brief discussion highlights an important point: that comparisons between HG and OT are affected by the way in which candidate sets are generated (see further Pater, 2009b).

Both gradient Alignment and the parallel evaluation of multiple instances of faithfulness constraint violation can be characterized as producing unwanted global effects in the standard version of OT. We thus have unwanted globality introduced by particular types of constraint, and by particular assumptions about how candidates are generated and evaluated. In both cases, it seems likely that refinements to the theory that impose the desired locality restrictions in OT will result in the elimination of many of the “complicated trade-offs” that may have dissuaded Prince and Smolensky from pursuing a version of OT with weighted constraints.

## 4. Learning and variation in HG

### 4.1. Emergent variation in Japanese loanwords

We now return to the Japanese data introduced in section 3.1. As well as completing the account of the data, this section briefly reviews and expands 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 noncategorical linguistic phenomena. HG was in fact originally motivated as an account of gradient syntactic well-formedness (Legendre et al., 1990; although cf. Legendre et al., 2006). The other argument is that weighted constraints are compatible with existing well-understood algorithms for learning variable outcomes and for learning gradually. As these algorithms are broadly applied with connectionist and statistical models of cognition, this forms an important connection between the HG version of GL and other research in cognitive science.

Up to this point, I 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. Brown, 2008; Coetzee & Pater, 2008b; Hayes & Wilson, 2008; Keller, 2006; Martin, 2007; McClelland & Vander Wyck, 2006). Finally, for each input, the choice of output is categorical in the sense that it cannot vary probabilistically (cf. Boersma & Pater, 2008; Goldrick & Daland, 2009; Goldwater & Johnson, 2003; Jäger & Rosenbach, 2006; Wilson, 2006). In the OT literature, probabilistic choice between outputs has been much applied as an account of phonological variation (see Anttila, 2007; Coetzee & Pater, 2008a for overviews).

I have also not addressed the question of how constraint weights might be set by a language learner. In comparison with some other theories of GL, one of the strengths of OT is that it has associated learning algorithms. The Constraint Demotion algorithms (Tesar & Smolensky, 1998, 2000) are guaranteed to find a ranking for any lan-

guage (set of optimal input–output pairs) that can be represented by a given set of OT constraints, if they are given full information about the structure of the language data. However, because the original version of OT for which Constraint Demotion is designed cannot represent variation between optima within a single candidate set, Constraint Demotion will fail to learn languages that display such variation (see, e.g., Boersma & Hayes, 2001).

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 2.1. As shown in (21), geminates are optionally devoiced when there is another voiced obstruent in the word (e.g., /dog:u/, [dok:u] *or* [dog:u]). The analysis in section 3.1 deals only 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 data that borrowers experienced for a difference between the forms that show devoicing and those that do not (e.g., /bobu/, [bobu] and /web:u/, [web:u], which lack the combination of a voiced geminate and another voiced obstruent).

There are two existing versions of HG that can produce variation between optima. In the probabilistic model of grammar proposed by Johnson (2000, 2002), a candidate's probability relative to the rest of the candidate set is proportional to the exponential of its harmony (often referred to as “Max-Ent-OT”; see also Goldwater & Johnson, 2003; Jäger, 2007; Wilson, 2006). In the other approach to stochastic HG, noise is placed on the constraint weights, producing a model we can call *Noisy HG* (see Boersma & Pater, 2008 for a formal presentation and further references; see also Goldrick & Daland, 2009 on the relationship to connectionist models of speech processing). A commonality between these models of grammar is that they can be learned by a simple on-line error-driven learning algorithm, referred to in the neural modeling and statistical learning literature as the perceptron convergence procedure (Rosenblatt, 1958), the delta rule (Widrow & Hoff, 1960), and stochastic gradient ascent/descent (in phonology see Jäger, 2007).<sup>11</sup> Error-driven learning of this type is, of course, broadly used in psycholinguistic modeling and has recently been claimed to be crucial to an account of the learning of syntactic generalization (Chang, Dell, & Bock, 2006). Boersma and Pater (2008) dub this algorithm HG-GLA (HG's Gradual Learning Algorithm); see that paper for a full description. For both models of stochastic HG, the learning procedure is convergent under the same conditions as for Constraint Demotion (see Fischer, 2005 and Boersma & Pater, 2008). The behavior of HG-GLA contrasts with Boersma's (1998) GLA for stochastic OT, which fails to converge on a simple nonvariable pattern (Pater, 2008a).

To illustrate how these models can deal with the Japanese data, I adopt a version of Noisy HG in which a constraint's violation score is multiplied by the exponent of the sum of that constraint's weight plus its random noise (Boersma & Pater, 2008). That is, harmony is now calculated as in (36), where  $N_k$  is a Gaussian random variable that is temporarily added to a constraint's weight when a candidate set is evaluated.

(36) *Noisy Exponential Harmony*

$$H = \sum_{k=1}^K s_k e^{w_k + N_k}$$

This exponentiation ensures that the violation score is always multiplied by a positive number, even if  $w_k + N_k$  happens to be negative. In this way, Exponential HG enforces the positivity restriction on weights that I assume throughout this paper.

We will now see how Noisy Exponential HG with HG-GLA can be used to model the incorporation of English loanwords into the phonology of Japanese. The simulation is performed in two phases. The first is intended to represent Japanese before the incorporation of English loanwords started to change its phonological system, that is, so-called native Japanese. To produce an appropriate grammar for this stage in the development of the language, we provide the learner with forms that would be typical of native Japanese. I simplify both the simulation and the presentation by using just the word types that are exemplified in the data in section 3.1, and by converting the loanwords to permissible native Japanese forms, as shown in (37). Rather than the loanword [bobu], we thus have [pobu], which respects the Lyman's Law restriction by devoicing the initial consonant. Rather than the loanword [web:u], we have [wep:u], which uses devoicing to conform to the restriction against voiced geminates. For [dok:u]/[dog:u], the learner is supplied with the version with devoicing of the geminate. I have also simplified the simulation by providing the learner with the (input, output) pairs, rather than just the observed outputs.

(37) *“Native Japanese” data*

/bobu/	[pobu]
/web:u/	[wep:u]
/dog:u/	[dok:u]

The learner is exposed to each of these data types in equal proportion. The learner's grammar is made up of the output constraints \*VOICE, \*VCE-GEM, and OCP-VOICE, with the faithfulness constraint IDENT-VOICE. Following Jesney and Tessier (2007), the output constraints begin with a greater weight (10) than the faithfulness constraint (0), although as I have provided (input, output) pairs, this is not crucial to the outcome here. The learning rate is set at 0.01, and the standard deviation of the noise is set at 0.2 (this simulation was conducted in Praat; Boersma & Weenink, 2008). After being presented with 100,000 input–output pairs, the learner's grammar was as in (38). The numbers above the constraints are the exponentiated weights; for example,  $w(*VOICE) = -3.04$ , which becomes 0.05 after exponentiation. The numbers to the left of the output candidates show the probability with which each one is selected by the grammar, estimated by sampling from 100,000 runs of the grammar with noise of 0.2. While the correct outputs are nearly always chosen, there is some residual devoicing of

obstruents that are voiced in the learning data; [popu] was chosen in 10 trials, and [tok:u] in 5. Candidates without numbers beside them were selected in none of the 100,000 trials (strictly speaking, no nonharmonically bounded candidate has zero probability in this theory).

(38) *Voicing in native Japanese phonology*

	12.18	12.18	0.14	0.05	
/bobu/	OCP-VOICE	*VCE-GEM	IDENT-VOICE	*VOICE	
bobu	-1			-2	-12.28
> 0.99 [pobu]			-1	-1	-0.19
< 0.01 [popu]			-2		-0.28
	12.18	12.18	0.14	0.05	
/web:u/	OCP-VOICE	*VCE-GEM	IDENT-VOICE	*VOICE	
web:u		-1		-1	-12.23
1 [wep:u]			-1		-0.14
	12.18	12.18	0.14	0.05	
/dog:u/	OCP-VOICE	*VCE-GEM	IDENT-VOICE	*VOICE	
dog:u	-1	-1		-2	-24.46
> 0.99 [dok:u]			-1	-1	-0.19
< 0.01 [tok:u]			-2		-0.28

In the second phase of the simulation, the learner is presented with words in their ‘‘English’’ form. Here, all of the obstruents are voiced, including the geminate of [dog:u]. Although English does not have voiced geminates, the corresponding singleton is voiced, and as the full set of loanword data in (20) above and the sample in (39) show, the voicing distinction is usually maintained in borrowings (although variably in the Lyman’s Law environment).

(39) *‘‘English’’ data*

/bobu/ [bobu]  
 /web:u/ [wep:u]  
 /dog:u/ [dog:u]

The parameters of the simulation otherwise remain the same as for the first phase. After the learner was presented with 400 input–output pairs randomly selected from those in (39), the grammar was as in (40). Now, the grammar chooses the faithful candidate 99% of the time for both /bobu/ and /web:u/, but varies between [dog:u] and [dok:u] for /dog:u/.

(40) Emergent variable Japanese loanword devoicing

	8.63	4.41	4.08	<0.01	
/bobu/	IDENT-VOICE	*V <sub>CE</sub> -GEM	OCP-VOICE	*V <sub>OICE</sub>	
> 0.99 [bobu]			-1	-2	-4.09
< 0.01 [pobu]	-1			-1	-8.63
popu	-2				-17.26

  

	8.63	4.41	4.08	<0.01	
/web:u/	IDENT-VOICE	*V <sub>CE</sub> -GEM	OCP-VOICE	*V <sub>OICE</sub>	
> 0.99 [web:u]		-1		-1	-4.42
< 0.01 [wep:u]	-1				-8.63

  

	8.63	4.41	4.08	<0.01	
/dog:u/	IDENT-VOICE	*V <sub>CE</sub> -GEM	OCP-VOICE	*V <sub>OICE</sub>	
0.51 [dog:u]		-1	-1	-2	-8.50
0.49 [dok:u]	-1			-1	-8.63
tok:u	-2				-17.26

The loanword pattern thus emerges from gradual learning (HG-GLA), and from stochastic evaluation (Noisy Exponential HG). This result also requires evaluation with weighted constraints. In stochastic OT (Boersma, 1998), these constraints will in fact produce a greater proportion of devoicing outcomes for /dog:u/ than for the other inputs. However, if the constraint values are such that the faithful outcome for /web:u/ and /bobu/ is nearly always chosen, there will be relatively little variation for /dog:u/.

This simulation is, of course, a gross abstraction from the actual process of loanword adaptation. A more realistic approach would incorporate both perception and production (see, e.g., Boersma & Hamann, 2008) and would model the cross-linguistic interaction of speakers and listeners (and even perhaps readers). It is adequate for present purposes in that it provides a demonstration that HG can be straightforwardly adapted to yield variable outcomes, and that the resulting theories are compatible with models of learning that have been developed outside of GL.

This approach to learning predicts that cumulative constraint interaction should be directly observable in language acquisition. This prediction appears to be correct. Working in OT, Pater and Werle (2003: Appendix) describe a pattern in one child’s productions that they analyze as a “cumulative effect of constraint violation, which cannot be captured in standard Optimality Theory with strict domination.” Between the ages of 1; 8 and 1; 11 (years; months), this child produced the words *keys*, *kiss*, and *cat* with an initial [t], rather than the initial [k] of the adult form. This change was conditioned by two factors. First, it only occurred if the adjacent vowel was in the *front* category, that is, a vowel that is produced with forward movement of the tongue within the oral cavity. Second, it only occurred if the following consonant was in the category of *coronals*, that is, a consonant that like [t] is produced with contact between the tongue tip and the oral cavity ([k] is a *dorsal* consonant, produced using the back of the tongue).

Neither condition was sufficient on its own. Converted to constraint violations, this would be identical to the Japanese loanword example: A faithfulness constraint (e.g., \*/K/ → [T], where K = dorsal and T = coronal) is violated only if this satisfies two output constraints (e.g., \*[KI] and \*[KVT], I = front vowel, V = vowel). This example suggests that application of HG to the study of phonological acquisition is a fruitful area for further research, as also shown by Jäger (2007), Jesney and Tessier (2007), Albright et al. (2008), and Farris-Trimble (2008).

#### 4.2. Constraint universality, learning, and typology

This section briefly addresses an issue that connects the study of language learning with the study of typology, and which may have been raised in the minds of many readers. The issue is the source of the constraints themselves.

One obvious interpretation of Prince and Smolensky's (1993/2004) proposal that constraints are universal is that they are innately available to the learner. However, there are of course reasons to explore the possibility that they are learned, and much research on learning in OT has in fact done so. Studies of human language acquisition in OT sometimes assume that constraints are constructed by the learner (e.g., Bernhardt & Stemberger, 1998; Pater, 1997). In Pater (1997), this assumption is driven by the observation that child language phonology contains patterns that occur in no adult language, with *consonant harmony* like that discussed in the last paragraph of the last section being the prime example (see Berg & Schade, 2000 for a connectionist processing account). Generative research on phonological acquisition has long grappled with divergences between child and adult phonology, and one response is to assume a shared formal system, with the differences lying in the substantive content of rules or constraints (Kiparsky & Menn, 1977; Pater, 2002; Smith, 1973).

Language-specific constraints are in fact proposed in Prince and Smolensky (1993/2004) to account for cases in which a phonological pattern is limited to particular words or morphemes: FREE-V in section 3.2 is an example of such a constraint. Pater (2008b) provides an overview of subsequent proposals and an OT learning account for one of them. Further, as many phonological constraints seem to be grounded in phonetics (i.e., articulation and perception), there have been several proposals in OT for induction of the constraints from phonetics (e.g., Flack, 2007; Hayes, 1999; Smith, 2002). Boersma (1998) proposes an alternative multilevel version of OT in which phonetic and phonological constraints are on separate levels and are both claimed to be learned from experience. Finally, working in a multilevel version of the present framework, Boersma and Pater (2007) sketch an account of how phonological constraints could be constructed based on phonological structures in the learning data (see Burzio, 2002; Hayes & Wilson, 2008 for related precedents), and also an account of how phonetic tendencies could become phonological constraints over the course of generations. As this approach to grammar learning assumes a set of representational primitives that are combined by the learner to form constraints, Boersma and Pater call this *innately guided empiricism*. It is related to what Goldwater (2007) calls *structured statistical learning*, and as she notes, this general framework holds much promise as a "middle way" between the extremes of highly nativist and highly empiricist theories of language learning. The pursuit of this line of research would make especially close contact

with current research elsewhere in cognitive science (see, e.g., Eisner's 2002 introduction and the papers in the special section on "linguistically apt statistical methods" in *Cognitive Science* 26).

As alluded to in the last paragraph, many phonological asymmetries have their source in phonetic factors. For example, section 2.1 focused on the asymmetry between the well-attested syllable-final devoicing pattern and the apparently unattested syllable-initial devoicing pattern. This asymmetry is almost certainly related to the relative difficulty of producing and perceiving voicing in final position (see Myers, 2008 for a review of the literature supporting this view). Given such a phonetic basis for a typological asymmetry, some have argued that phonological theory should not attempt to account for that asymmetry in terms of a universal property of phonological representations (e.g., Ohala, 1990), or of constraint sets (e.g., Hale & Reiss, 2000). However, one can certainly maintain constraint universality as a useful heuristic for the study of typology (see Ellison, 2000 for arguments for this sort of pragmatic approach to universality) and remain agnostic about whether the ultimate explanation for typological asymmetries lies in language use, language learning, the innate structure of cognitive-linguistic representations, or as is most likely, some combination of all of these. The elaboration of such explanations is as an interdisciplinary enterprise, which may well be aided by the adoption of a framework for GL that is closely connected to models in other disciplines. If typological research is conducted with weighted rather than ranked constraints, the results will probably be more easily brought to bear on questions such as how much prior knowledge must be given to a learning algorithm for it to acquire the full range of human languages (see again Hayes & Wilson, 2008 in phonology), what the relationships are between typologically observed patterns and observed biases in language learning and processing (see, e.g., Moreton, 2008), and how theories of these relationships can be fully formalized and tested.

## 5. Conclusions

In this paper, I have argued for the adoption of HG as a framework for GL. As discussed in section 4, there is a body of earlier work that provides motivation for the replacement of OT's ranked constraints with weighted ones. Here I have addressed a set of influential claims, reviewed in section 1, that this would result in a theory that fails to meet a core goal of GL: restrictiveness in typological predictions. Insofar as the examples of symmetric trade-offs in section 2 and of vacuous gang effects in section 3.2 are typical of many cases of constraint interaction in OT, I have shown that many of OT's typological results translate directly into a version of the theory with positively weighted constraints. I have also provided an example of how HG's weighted constraints allow it to capture attested linguistic patterns that escape OT rankings of the same constraints (section 3.1). As theories of typology in HG can, therefore, use different constraint sets than ones in OT, comparing the predictions of the two frameworks will require a great deal of cross-linguistic analysis. A likely reason that the typological predictions of HG have been so little studied, besides the

influence of the claims that they are obviously wrong, is that linguistic analysis can be difficult with numerically weighted constraints. This barrier can be overcome by the use of computational techniques, as shown by Potts et al. (2009).

It is important to emphasize that in the small set of results currently available on the comparison of OT and HG typologies one can also find vindication for the view that HG does generate implausible patterns not generated by OT. In section 3.3, I suggested alternative diagnoses for the sources of some of these pathologies: that they result from assumptions about the nature of constraints and about the nature of candidate generation and evaluation that are controversial even within OT (and/or from the relative difficulty of learning the unattested systems). The fleshing out of versions of HG and OT that adopt alternative assumptions in these regards is an ongoing area of research. The results of this work should help to determine whether strict domination is in fact crucial to the success of an optimization-based approach to GL. As discussed in section 4.2, a separate line of ongoing research seeks to explain typological generalizations without assuming that the universal constraint set is a cognitive universal. This research should also shed light on whether strict domination is required for an account of typology.

Much of the appeal of HG comes from its compatibility with models of cognition, and especially of learning, developed outside of GL. There is a great deal of potential for interdisciplinary research on language learning and other types of processing in HG; section 4 noted the potential in this regard for the study of phonological acquisition. HG also occupies a central place in Smolensky and Legendre's (2006) integrated connectionist/symbolic theory of mind. However, their view of HG is quite different from the one being argued for here. Legendre et al. (2006) state that "HG is not and was never intended to be a theory of typology" (p. 373), and they see it as a sort of bridge to the fully symbolic grammatical module of OT. If HG can in fact replace OT in the modeling of linguistic knowledge and typology, then a dramatic simplification of Smolensky and Legendre's (2006) theory might well result.

## Notes

1. In many languages with final devoicing, including Russian, coda obstruents can be voiced when there is a voiced obstruent in the beginning of the following syllable. See Lombardi (1999) for an analysis in terms of constraint interaction.
2. It has been shown that final devoicing is in some cases not categorical, yielding an outcome between a fully voiced or voiceless consonant (Port & Crawford, 1989). See Flemming (2001) for a version of HG that can handle such between-category outcomes. It is also worth noting that it is possible to reanalyze "devoicing" as word-specific intervocalic voicing (Becker, Nevins, & Ketrez, 2008). And finally, one might avoid transformations altogether and use the constraints to choose between lexically stored alternatives (see McCarthy, 2002 for an overview of *allomorphy* in OT). I follow the traditional phonological analysis for ease of exposition and theory comparison.

3. An alternative declarative definition involves pairwise relations between candidates: When compared with any other candidate, the optimum better satisfies the highest ranked constraint that distinguishes between the two.
4. Adding \*VOICE to the constraint set changes the weighting condition for coda devoicing; see section 3.2.
5. Prince (2002) shows with an abstract set of violation profiles that OT collective harmonically bounding is not guaranteed to survive the translation into HG, and Prince (2007b) and Tesar (2007) reiterate the point. However, when violations trade off one to one as in (14), collective bounding does continue to hold in HG.
6. Elsewhere in the paper, I use the faithfulness constraints of McCarthy and Prince (1999), who replace PARSE with MAX, which directly penalizes deletion. I retain PARSE so that I do not have to alter Prince and Smolensky's examples.
7. The other case mentioned in this passage, that of the number of consonants in a cluster, also involves a symmetric trade-off if every consonant in a coda violates NoCODA.
8. Prince (2002) provides a formal characterization of the conditions under which one can replace the ranking relation  $\gg$  of an OT grammar with the numerical greater than relation  $>$ , assign to the constraints *any* set of weights that satisfies the demands of the resulting inequalities, and obtain an HG grammar that picks the same optima as the original OT one, that is, the conditions under which "Anything Goes." Prince's concern is the formal characterization of Anything Goes systems, rather than the issue of the typological predictions of HG and OT. Although he discusses the equivalence of OT and HG under symmetric trade-offs, he does not discuss the consequences for Prince and Smolensky's (1997) argument for strict domination. Systems that are not Anything Goes can produce identical typological results in HG and OT (see, e.g., section 3.2), and where they do not, it can, of course, be the case that HG produces the correct result (see, e.g., section 3.1).
9. In terms of the present model, a restriction on the size of windows can be obtained by imposing a maximum value on the weights and a minimum difference between the harmony scores of the optima and their competitors (see Boersma & Pater, 2008 and Potts et al., 2009 on the margin of separation of harmony; see relatedly Albright et al., 2008). For example, if we require the optimal candidate's harmony to exceed that of any of its competitors by at least 0.5, then a three-syllable window requires a minimum constraint weight of 2.5, while a four-syllable window requires 3.5. The requirements of a large margin of separation and a maximum value on weights are commonly imposed in learning algorithms in the machine learning and neural modeling literature, and seem quite plausible as biological limitations (in terms of the learning model in section 4.1, a minimum margin of separation is needed to overcome a given amount of noise).
10. A third source is in trade-offs involving faithfulness constraints whose satisfaction can require an unbounded number of violations of output constraints, or of other faithfulness constraints. There is maybe only one such constraint: the one violated by the Null

Parse, which Prince and Smolensky (1993/2004: 57–61) dub  $L_X \approx PR$ . The consequences of the trade-offs with this relatively infrequently used constraint deserve further investigation.

11. There may in fact be a (much) more fundamental connection between noisy and probabilistic versions of HG, in that it has been proposed that neural noise is crucial to the human computation of probabilities (see Salinas, 2006 for an accessible summary).

## Acknowledgments

I especially thank Rajesh Bhatt and Christopher Potts for their collaboration on earlier presentations of some of this material, and the reviewers and associate editor for their helpful advice. The development of this work was also particularly helped by discussions with Adam Albright, Michael Becker, Paul Boersma, Andries Coetzee, Jason Eisner, Robert Frank, Matt Goldrick, John Goldsmith, Bruce Hayes, Karen Jesney, Mark Johnson, Shigeto Kawahara, René Kager, John McCarthy, Elliott Moreton, Patrick Pratt, Alan Prince, Kathryn Pruitt, Jason Riggle, Lisa Selkirk, Brian Smith, Paul Smolensky, Bruce Tesar, Anne-Michelle Tessier, and Colin Wilson, as well with other participants in courses and conferences where parts of it were presented. This research was supported by a Faculty Research Grant from the University of Massachusetts, Amherst, and by grant BCS-0813829 from the National Science Foundation to the University of Massachusetts, Amherst.

## References

- Albright, A., Magri, G., & Michaels, J. (2008). Modeling doubly marked lags with a split additive model. In H. Chan, H. Jacob, & E. Kiparsky (Eds.), *BUCLD 32: Proceedings of the 32nd annual Boston University conference on language development* (pp. 36–47). Somerville, MA: Cascadia Press.
- Anttila, A. (2007). Variation and optionality. In P. de Lacy (Ed.), *The Cambridge handbook of phonology* (pp. 519–536). Cambridge, England: Cambridge University Press.
- Barlow, J. A., & Gierut, J. A. (1999). Optimality theory in phonological acquisition. *Journal of Speech, Language, and Hearing Research*, 42, 1482–1498.
- Becker, M., Nevins, A., & Ketrez, N. (2008). The surfeit of the stimulus: Grammatical biases filter lexical statistics in Turkish voicing deneutralization. Available at: <http://roa.rutgers.edu/>. Accessed on May 26, 2009.
- Becker, M., Pater, J., & Potts, C. (2007). OT-Help: Java tools for Optimality Theory. Software available online at: <http://web.linguist.umass.edu/~OTHelp/>.
- Berg, T., & Schade, U. (2000). A local connectionist account of consonant harmony in child language. *Cognitive Science*, 24, 123–149.
- Bernhardt, B. H., & Stemberger, J. P. (1998). *Handbook of phonological development from the perspective of constraint-based nonlinear phonology*. San Diego, CA: Academic Press.
- Bíró, T. (2003). Quadratic alignment constraints and finite state optimality theory. In *Proceedings of the workshop on finite-state methods in natural language processing* (pp. 119–126). Budapest, Hungary. Available at: <http://roa.rutgers.edu/>. Accessed on May 26, 2009.

- Boersma, P. (1998). *Functional phonology: Formalizing the interaction between articulatory and perceptual drives*. The Hague, The Netherlands: Holland Academic Graphics.
- Boersma, P., & Hamann, S. (2008). Loanword adaptation as first-language phonological perception. Available at: <http://roa.rutgers.edu/>. Accessed on May 26, 2009.
- Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32, 45–86. Available at: <http://roa.rutgers.edu/>.
- Boersma, P., & Levelt, C. (2003). Optimality Theory and phonological acquisition. *Annual Review of Language Acquisition*, 3, 1–50.
- Boersma, P., & Pater, J. (2007). Constructing constraints from language data: The case of Canadian English diphthongs. Paper presented to the North Eastern Linguistic Society. Handout available at: <http://people.umass.edu/pater/boersma-pater-nels.pdf>.
- Boersma, P., & Pater, J. (2008). Convergence properties of a gradual learning algorithm for Harmonic Grammar. Available at: <http://roa.rutgers.edu/>. Accessed on May 29, 2009.
- Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer (Version 5.0.17) [Computer program]. Retrieved April 1, 2008 from <http://www.praat.org/>. Developed at the Institute of Phonetic Sciences, University of Amsterdam.
- Brown, J. (2008). *Theoretical aspects of Gitskan phonology*. PhD thesis, University of British Columbia.
- Burzio, L. (2002). Missing players: Phonology and the past-tense debate. *Lingua*, 112, 157–199.
- Bybee, J. (2001). *Phonology and language use*. Cambridge, England: Cambridge University Press.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113(2), 234–272.
- Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.
- Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, The Netherlands: Foris.
- Coetzee, A., & Pater, J. (2008a). The place of variation in phonological theory. In J. Goldsmith, J. Riggle, & A. Yu (Eds.), *The handbook of phonological theory* (2nd edn.). Malden, MA: Blackwell. Available at: <http://roa.rutgers.edu/>.
- Coetzee, A., & Pater, J. (2008b). Weighted constraints and gradient phonotactics in Muna and Arabic. *Natural Language and Linguistic Theory*, 26, 289–337.
- Comrie, B. (1976). Irregular stress in Polish and Macedonian. *International Review of Slavic Linguistics*, 1, 227–240.
- Dresher, E. (1994). Acquiring stress systems. In E. S. Ristad (Ed.), *Language computations, number 17 in DIMACS series in discrete mathematics and theoretical computer science* (pp. 71–92). Providence, RI: AMS.
- Eisner, J. (1998). FootForm decomposed: Using primitive constraints in OT. In B. Bruening (Ed.), *Proceedings of SCIL VIII*, number 31 in MIT Working Papers in Linguistics (pp. 115–143). Cambridge, MA: Department of Linguistics, MIT.
- Eisner, J. (2002). Introduction to the special section on linguistically apt statistical methods. *Cognitive Science*, 26, 235–237.
- Ellison, T. M. (2000). The universal constraint set: Convention, not fact. In J. Dekkers, F. van der Leeuw, & J. van der Weijer (Eds.), *Optimality theory: Phonology, syntax, and acquisition* (pp. 524–553). Oxford, England: Oxford University Press.
- Everett, D., & Everett, K. (1984). On the relevance of syllable onsets to stress placement. *Linguistic Inquiry*, 15, 705–711.
- Farris-Trimble, A. (2008). *Cumulative faithfulness effects in phonology*. PhD thesis, Indiana University.
- Fischer, M. (2005). *A Robbins-Monro type learning algorithm for a maximum-entropy maximizing version of stochastic Optimality Theory*. Master's thesis, Humboldt University, Berlin. Available at: <http://roa.rutgers.edu/>.
- Flack, K. (2007). *The sources of phonological markedness*. PhD thesis, University of Massachusetts, Amherst, MA.
- Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology*, 18, 7–44.

- Goldrick, M., & Daland, R. (2009). Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology*, 26(1), 147–185.
- Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Oxford, England: Blackwell.
- Goldsmith, J. (1991). Phonology as an intelligent system. In D. J. Napoli & J. Kegl (Eds.), *Bridges between psychology and linguistics: A swarthmore festschrift for lila gleitman* (pp. 247–267). Hillsdale, NJ: Erlbaum.
- Goldsmith, J. (1993a). Harmonic phonology. In J. Goldsmith (Ed.), *The last phonological rule* (pp. 21–60). Chicago: University of Chicago Press.
- Goldsmith, J. (1993b). *The last phonological rule: Reflections on constraints and derivations*. Chicago: University of Chicago Press.
- Goldsmith, J. (1994). A dynamic computational theory of accent systems. In J. Cole & C. Kisseberth (Eds.), *Perspectives in phonology* (pp. 1–28). Stanford, CA: CSLI.
- Goldwater, S. (2007). *Nonparametric Bayesian models of lexical acquisition*. PhD thesis, Brown University.
- Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, & O. Dahl (Eds.), *Proceedings of the Stockholm workshop on variation within optimality theory* (pp. 111–120). Stockholm: Stockholm University.
- Gordon, M. (2002). A factorial typology of quantity-insensitive stress. *Natural Language and Linguistic Theory*, 20, 491–552.
- Gupta, P., & Touretzky, D. S. (1994). Connectionist models and linguistic theory: Investigations of stress systems in language. *Cognitive Science*, 18(1), 1–50.
- Guy, G. R. (1997). Violable is variable: Optimality Theory and linguistic variation. *Language Variation and Change*, 9, 333–347.
- Hale, K. (1973). Deep-surface canonical disparities in relation to analysis and change: An Australian example. In T. Sebeok (Ed.), *Current trends in linguistics, volume 9: Diachronic, areal and typological linguistics* (pp. 401–458). The Hague, The Netherlands: Mouton.
- Hale, M., & Reiss, C. (2000). Substance abuse and ‘dysfunctionalism’: Current trends in phonology. *Linguistic Inquiry*, 31, 157–169.
- Hammond, M. (1991). Parameters of metrical theory and learnability. In I. Roca (Ed.), *Logical issues in language acquisition* (pp. 47–62). Dordrecht, The Netherlands: Foris.
- Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford, England: Oxford University Press.
- Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: The University of Chicago Press.
- Hayes, B. (1999). Phonetically driven phonology: The role of Optimality Theory and inductive grounding. In M. Darnell, F. J. Newmeyer, M. Noonan, E. Moravcsik, & K. Wheatley (Eds.), *Functionalism and formalism in linguistics, volume 1: General papers* (pp. 243–285). Amsterdam: John Benjamins.
- Hayes, B., Tesar, B., & Zuraw, K. (2003). OTSoft 2.1. Software package developed at UCLA.
- Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39, 379–440.
- Hume, E. (2001). Metathesis: Formal and functional considerations. In E. Hume, N. Smith, & J. van de Weijer (Eds.), *Surface syllable structure and segment sequencing* (pp. 1–25). Leiden, The Netherlands: HIL.
- Hyde, B. (2007). An alignment approach to stress windows. Colloquium presentation, Rutgers University. Available at: <http://www.artsci.wustl.edu/bhyde/windowshandout.pdf>.
- Ito, J., & Mester, A. (1986). The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry*, 17, 49–73.
- Ito, J., & Mester, A. (1995). Japanese phonology. In J. Goldsmith (Ed.), *Handbook of phonological theory* (pp. 817–838). Cambridge, MA: Blackwell.
- Ito, J., & Mester, A. (2003). *Japanese morphophonemics: Markedness and word structure*. MIT Press Linguistic Inquiry Monograph Series 41. Cambridge, MA: MIT Press.
- Jäger, G. (2007). Maximum entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, & A. Zaenen (Eds.), *Architectures, rules, and preferences: A Festschrift for Joan Bresnan* (pp. 467–479). Stanford, CA: CSLI.

- Jäger, G., & Rosenbach, A. (2006). The winner takes it all—almost. Cumulativity in grammatical variation. *Linguistics*, 44, 937–971.
- Jesney, K., & Tessier, A.-M. (2007). Re-evaluating learning biases in Harmonic Grammar. In M. Becker (Ed.), *University of Massachusetts occasional papers in linguistics 36: Papers in theoretical and computational phonology* (pp. 69–110). Amherst, MA: Department of Linguistics, University of Massachusetts, Amherst; Graduate Linguistic Student Association.
- Johnson, M. (2000). Stochastic lexical-functional grammars. Paper presented to the LFG 2000 Conference. Slides available at: <http://www.cog.brown.edu/mj/papers/lfgOO-slides.pdf>.
- Johnson, M. (2002). Optimality-theoretic lexical functional grammar. In S. Stevenson & P. Merlo (Eds.), *The lexical basis of syntactic processing: Formal, computational and experimental issues* (pp. 59–73). Amsterdam: John Benjamins.
- Kager, R. (1999). *Optimality theory*. Cambridge, England: Cambridge University Press.
- Kager, R., Pater, J., & Zonneveld, W. (2004). *Constraints in phonological acquisition*. Cambridge, England: Cambridge University Press.
- 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 G. Fanselow, C. Féry, R. Vogel, & M. Schlesewsky (Eds.), *Gradience in grammar: Generative perspectives* (pp. 270–287). Oxford, England: Oxford University Press.
- Kiparsky, P., & Menn, L. (1977). On the acquisition of phonology. In J. Macnamara & J. Macnamara (Eds.), *Language learning and thought* (pp. 47–78). New York: Academic Press.
- de Lacy, P. (2006). *Markedness: Reduction and preservation in phonology*. Cambridge Studies in Linguistics 112. Cambridge, England: Cambridge University Press.
- Legendre, G., Miyata, Y., & Smolensky, P. (1990). Can connectionism contribute to syntax? Harmonic Grammar, with an application. In M. Ziolkowski, M. Noske, & K. Deaton (Eds.), *Proceedings of the 26th regional meeting of the Chicago Linguistic Society* (pp. 237–252). Chicago: Chicago Linguistic Society.
- Legendre, G., Sorace, A., & Smolensky, P. (2006). The Optimality Theory-Harmonic Grammar connection. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: From neural computation to optimality theoretic grammar, vol. 2: Linguistic and philosophical implications* (pp. 339–402). Cambridge, MA: MIT Press.
- Lombardi, L. (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *Natural Language and Linguistic Theory*, 17, 267–302.
- Lyman, B. S. (1894). Change from surd to sonant in Japanese compounds. *Oriental Studies of the Oriental Club of Philadelphia*, 1–17.
- Martin, A. (2007). *The evolving lexicon*. PhD thesis, UCLA.
- McCarthy, J. J. (2002). *A thematic guide to optimality theory*. Cambridge, England: Cambridge University Press.
- McCarthy, J. J. (2003). OT constraints are categorical. *Phonology*, 20(1), 75–138.
- McCarthy, J. J. (Ed.) (2004). *Optimality theory in phonology: A reader*. Malden, MA: Blackwell.
- McCarthy, J. J. (2007). Restraint of analysis. In S. Blaho, P. Bye, & M. Kramer (Eds.), *Freedom of analysis* (pp. 203–231). Berlin: Mouton de Gruyter.
- McCarthy, J. J., & Prince, A. (1993). Generalized Alignment. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology* (pp. 79–153). Dordrecht, The Netherlands: Kluwer. Excerpts appear in J. Goldsmith (Ed.) (1999), *Essential readings in phonology* (pp. 102–136). Oxford, England: Blackwell.
- McCarthy, J. J., & Prince, A. (1999). Faithfulness and identity in Prosodic Morphology. In R. Kager, H. van der Hulst, & W. Zonneveld (Eds.), *The prosody–morphology interface* (pp. 218–309). Cambridge, England: Cambridge University Press.
- McClelland, J. L., & Vander Wyck, B. (2006). Graded constraints in English word forms. Available at: <http://www-psych.stanford.edu/jlm/papers/>. Accessed on May 26, 2009.
- Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, 25(1), 83–127.

- Myers, S. (2008). *Final devoicing: An experimental investigation*. Austin, TX: University of Texas.
- Newmeyer, F. J. (2005). *Possible and probable languages*. Oxford, England: Oxford University Press.
- Nishimura, K. (2003). *Lyman's Law in loanwords*. Master's thesis, Nagoya University.
- Nishimura, K. (2006). Lyman's Law in loanwords. *Phonological Studies [Onin Kenkyuu]*, 9, 83–90.
- Ohala, J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology, vol. 1* (pp. 258–275). Cambridge, England: Cambridge University Press.
- Pater, J. (1997). Minimal violation and phonological development. *Language Acquisition*, 6, 201–253.
- Pater, J. (2002). Form and substance in phonological development. In L. Mikkelsen & C. Potts (Eds.), *WCCFL 21 Proceedings* (pp. 348–372), Somerville, MA: Cascadilla Press.
- Pater, J. (2008a). Gradual learning and convergence. *Linguistic Inquiry*, 39(2), 334–345.
- Pater, J. (2008b). Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In S. Parker (Ed.), *Phonological argumentation: Essays on evidence and motivation*. London: Equinox. Available at: <http://roa.rutgers.edu/>.
- Pater, J. (2009a). Review of "The harmonic mind: From neural computation to optimality theoretic grammar," *Phonology*, 26(1), 217–226. Available at: <http://roa.rutgers.edu/>.
- Pater, J. (2009b). Serial harmonic grammar and Berber syllabification. In T. Borowsky, S. Kawahara, T. Shinya, & M. Sugahara (Eds.), *Prosody matters: Essays in honor of Elisabeth O. Selkirk*, England: Equinox Press.
- Pater, J., & Werle, A. (2003). Direction of assimilation in child consonant harmony. *Canadian Journal of Linguistics*, 48, 385–408.
- Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28, 73–193.
- Port, R., & Crawford, P. (1989). Incomplete neutralization and pragmatics in German. *Journal of Phonetics*, 17, 257–282.
- Potts, C., Pater, J., Jesney, K., Bhatt, R., & Becker, M. (2009). Harmonic Grammar with linear programming: From linear systems to linguistic typology. Available at: <http://roa.rutgers.edu/>. Accessed on May 29, 2009
- Potts, C. and Pullum, G. K. (2002). Model theory and the content of OT constraints. *Phonology*, 19(3), 361–393.
- Prince, A. (1990). Quantitative consequences of rhythmic organization. In M. Ziolkowski, M. Noske, & K. Deaton (Eds.), *Parasession on the syllable in phonetics and phonology* (pp. 355–398). Chicago: Chicago Linguistic Society.
- Prince, A. (1993). In defense of the number i: Anatomy of a linear dynamical model of linguistic generalizations. RuCCS Technical Report 1. New Brunswick, NJ: Rutgers University.
- Prince, A. (2002). Anything goes. In T. Honma, M. Okazaki, T. Tabata, & S. ichi Tanaka (Eds.), *New century of phonology and phonological theory* (pp. 66–90). Tokyo: Kaitakusha. Available at: <http://roa.rutgers.edu/>.
- Prince, A. (2007a). In pursuit of theory. In P. de Lacy (Ed.), *Cambridge handbook of phonology* (pp. 33–60). Cambridge, England: Cambridge University Press.
- Prince, A. (2007b). Let the decimal system do it for you: A ridiculously simple utility function for OT. Available at: <http://roa.rutgers.edu/>. Accessed May 26, 2009.
- Prince, A., & Smolensky, P. (1993/2004). Optimality Theory: Constraint interaction in generative grammar. RuCCS Technical Report 2. Piscataway, NJ: Rutgers University Center for Cognitive Science, Rutgers University. Revised version published 2004 by Blackwell. (Page references in the text are to the 2004 version.)
- Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, 275, 1604–1610.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408.
- Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart, t. P. R. Group, J. L. McClelland, D. E. Rumelhart, & t. P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press/Bradford Books.
- Salinas, E. (2006). Noisy neurons can certainly compute. *Nature Neuroscience*, 9(11), 1349–1350.

- Samek-Lodovici, V., & Prince, A. (1999). Optima. RuCCS Technical Report no. 57, Piscataway, NJ: Rutgers University Center for Cognitive Science. Available at: <http://roa.rutgers.edu>. Accessed on May 26, 2009.
- Smith, N. V. (1973). *The acquisition of phonology: A case study*. Cambridge, England: Cambridge University Press.
- Smith, J. (2002). *Phonological augmentation in prominent positions*. PhD thesis, UMass Amherst, Amherst, MA.
- Smolensky, P. (2006a). Harmony in linguistic cognition. *Cognitive Science*, 30, 779–801.
- Smolensky, P. (2006b). Optimality in phonology II: Harmonic completeness, local constraint conjunction, and feature domain markedness. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: From neural computation to optimality theoretic grammar, vol. 2: Linguistic and philosophical implications* (pp. 27–160). Cambridge, MA: MIT Press.
- Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality theoretic grammar*. Cambridge, MA: MIT Press.
- Tesar, B. (2007). A comparison of lexicographic and linear numeric optimization using violation difference ratios. Available at: <http://roa.rutgers.edu/>. Accessed on May 26, 2009.
- Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, 29, 229–268.
- Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory*. Cambridge, MA: MIT Press.
- Widrow, G., & Hoff, M. (1960). Adaptive switching circuits. In *Western electronic show and convention record* (pp. 96–104). Institute of Radio Engineers. (Reprinted in J. Anderson and E. Rosenfeld. 1988. *Neurocomputing: Foundations of research*. Cambridge, MA: MIT Press, pp. 123–134.)
- Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5), 945–982.