Harmonic Grammar, Gradual Learning, and Phonological Gradience

Joe Pater, University of Massachusetts, Amherst

Stanford Workshop on Variation, Gradience and Frequency in Phonology, July 7th 2007

In proposing constraint ranking Prince and Smolensky (1993/2004) depart from OT's predecessor Harmonic Grammar (HG; Smolensky and Legendre 2006). In HG optimality is defined numerically as maximal Harmony, which is calculated as the sum of a candidate's weighted constraint scores.

OT's use of ranking rather than weighting is sometimes justified in terms of its restrictiveness: weighted interaction is claimed to be too powerful for HG to function as a realistic model of human language (Prince and Smolensky 1993/2004: 236; 1997: 1608, Legendre et al. 2006b).

In this talk, I present results of several ongoing collaborative research projects on HG. I will discuss the following points:

(1) i. HG is (perhaps surprisingly) restrictive, due to inherent limitations on the types of languages that can be generated by an optimization system (Bhatt et al. 2007; Pater et al. 2007)

ii. HG is compatible with a simple correctly convergent gradual learning algorithm, the Perceptron algorithm of Rosenblatt (1958) (Boersma and Pater 2007; Pater 2007a; see Jäger 2006, Soderstrom et al. 2006 for precedents).

iii. To deal with variation, HG can be implemented with noise, as in stochastic OT (Boersma 1998; Boersma and Hayes 2001). Testing shows that noisy HG+Perceptron is robust, unlike stochastic OT+GLA (Boersma and Pater 2007).

iv. Gradual learning yields Harmony values that reflect frequency distributions. A problem for the HG account of gradient well-formedness (Keller 2006; Legendre et al. 2006a) raised by Boersma (2004) can be resolved with a revised HG acceptability metric (Coetzee and Pater 2007).

1 Weighted optimization

Given a linguistic representation R, its scores on a set of constraints ($\{C_1(R), C_2(R), C_3(R),...C_n(R)\}$), and a set of coefficients, or weights for the constraints ($\{W_1, W_2, W_3,...W_n\}$), HG's Harmony function returns the sum of the weighted constraint scores.

(2) $\mathcal{H}(R) = (C_1(R)*W_1) + (C_2(R)*W_2) + (C_3(R)*W_3) +... (C_n*W_n)$

As Prince and Smolensky (1993/2004: 236) point out, optimality in an OT-like theory of generative grammar can be defined in terms of Harmony (see also Keller 2000, 2006; Flemming 2001; Prince 2002; Legendre et al.

2006b; Soderstrom et al. 2006). In the following OT-style tableau, the top row provides the constraint weights, and the rightmost column provides the Harmony score of each candidate. Constraint violations are indicated as negative integers, and the optimal candidate has the maximal Harmony (Legendre et al. 2006b):

(3) A weighted constraint tableau

	Weight	1.5	1	$\mathcal{H}$
	/Input-1/	CONSTRAINT-1	CONSTRAINT-2	
☞	Output-11		-1	-1
	Output-12	-1		-1.5

Final consonant devoicing in HG:

(4) Coda devoicing

	Weight	1.5	1	$\mathcal{H}$
	/bad/	*CODA-VOICE	IDENT-VOICE	
☞	bat		-1	-1
	bad	-1		-1.5

In this situation, weighting functions just like ranking. So long as the weight of *CODA-VOICE is greater than that of IDENT-VOICE, coda devoicing is produced. If IDENT-VOICE has a greater weight than *CODA-VOICE, voiced codas are allowed:

(5) w(*CODA-VOICE) $>$ w(IDENT-VOICE) = *CODA-VOICE $\gg$ IDENT-VOICE
w(IDENT-VOICE) $>$ w(*CODA-VOICE) = IDENT-VOICE $\gg$ *CODA-VOICE

To illustrate the main difference between HG and OT, we can consider a slightly more complicated example in the phonology of voicing.

In Japanese loanwords (but not native words), pairs of voiced obstruents are permitted (Nishimura2003, Kawahara 2006; all data are from the latter source):

(6) Violations of Lyman's Law in loanwords

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

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

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

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

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

(8) Optional devoicing of a geminate in Lyman's Law environment

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

According to Nishimura (2003) and Kawahara (2006), such devoicing is judged unacceptable in both (6) and (7).

An HG analysis:

(9) Japanese loanword devoicing as cumulative constraint interaction

Weight	1.5	1	$\mathcal{H}$
/bobu/	IDENT-VOICE	*2-VOICE	
☞ [bobu]		-1	-1
[bopu]	-1		-1.5

Weight	1.5	1	$\mathcal{H}$
/webbu/	IDENT-VOICE	*VCE-GEM	
☞ [webbu]		-1	-1
[weppu]	-1		-1.5

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

Due to the strict domination property of ranked constraints, which rules out cumulativity, or gang effects, no OT ranking of these constraints can generate this pattern (see Nishimura 2003 and Kawahara 2006 for expansions of the OT constraint set that deal with these data).

1.1 Optimization restricts cumulativity

Optimizing HG does not generate many of the cumulative interactions that one might expect *a priori*. For example, HG is incapable of generating a one-coda maximum (cf. Prince and Smolensky 1997: 1606; see also Prince 2002):

(10) No one-coda maximum in HG

/bat/	NOCODA	MAX
☞ bat	-1	
ba		-1

/bantat/	NoCoda	Max
bantat	-2	
☞ banta	-1	-1

Because the *weighting conditions* required for the optima in (10) are contradictory, no HG grammar can choose them:

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

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

The difference in the Japanese example above:

(12) A single Ident-Voice violation results in the satisfaction of both *2-Vce and *Vce-Gem

In the NoCoda/Max example, whether each potential coda is parsed is an independent decision, and it is therefore impossible to get a cumulative effect. This restrictiveness result would not hold with a numerical cut-off, or threshold, on well-formedness:

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

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

Because of the inherent restrictiveness of optimization, HG does not generate the unattested cases of cumulative constraint interaction identified in critiques of Smolensky's (2006) OT with Local Constraint Conjunction (see esp. McCarthy 1999, 2003a); see Pater (2006); Pater et al. (2007) for further discussion.

1.2 Locality and the power of HG

Given a set of constraints, HG is more powerful than OT:

(14) Any linguistic system that can be analyzed with a given set of ranked constraints can also be analyzed with weighted ones (Prince and Smolensky, 1993/2004: 236; Prince, 2002).

While the greater power of HG has been claimed to be excessive, there appears to be only one example in the published literature of unattested systems produced by weighting but not ranking:

(15) Weightings of Align-Head and Weight-to-Stress yield languages in which a heavy syllable is stressed if it is n syllables from the edge, but not if is is n+1 syllables away, where n is any integer (Legendre et al., 2006b)

It is clear that with gradient Alignment, HG overgenerates. The HG figures in (16) were obtained by submitting OT-Soft (Hayes et al. 2003) files prepared by René Kager to HaLP (Potts et al. 2007), which uses linear programming to find feasible combinations of optima (Bhatt et al. 2007). Positive HG limits constraint weights to positive reals (Prince 2002; Pater et al. 2007; Boersma and Weenink 2007; cf. Keller's 2000; 2006 limitation to non-negative reals). The OT figures were obtained by using Michael Becker's implementation of Recursive Constraint Demotion (Tesar and Smolensky 1998), which runs alongside HaLP in OT-Help (Becker et al. 2007).

(16) *Number of predicted languages with Kager's (2001) gradient Alignment constraint set*

All logically possible combinations of optima (words 2-9 syllables in length, candidates are all footings with QI binary trochees): 685, 292, 000

Optimality Theory: 35

Positive Harmonic Grammar: 911

It is less clear that HG overgenerates when independently motivated locality restrictions are imposed on the theory (Pater et al., 2007). For example, if gradient Alignment constraints are eliminated (see esp. McCarthy 2003b), the HG typology is reduced dramatically:

(17) *Number of predicted languages with Kager's (2001) non-gradient Alignment + Lapse constraints*

All logically possible combinations of optima: 685, 292, 000

Optimality Theory: 25

Positive Harmonic Grammar: 85

The HG typology would likely shrink further if the evaluation of representations by constraints was made even more local, in the senses discussed in Pater et al. (2007). Furthermore, a constraint set suitable for HG may be different from one that is suitable for OT (Pater 2007b).

2 Gradual learning in HG

One advantage of HG over OT is that it is compatible with the broad range of learning algorithms (and estimation and optimization procedures) that have been developed for linear and log-linear models of grammar (Johnson 2002; Goldwater and Johnson 2003; Hayes and Wilson 2006; Jäger 2006; Soderstrom et al. 2006; Wilson 2006).

A particularly simple such algorithm is the Perceptron update rule of Rosenblatt (1958). Like the Gradual Learning Algorithm for stochastic OT (GLA; Boersma 1998; Boersma and Hayes 2001), and some implementations of the constraint demotion algorithm (Tesar and Smolensky 1998, 2000), Perceptron is on-line and error-driven. Given the learner's error and the correct observed mapping, the weight of each constraint is updated as follows:

(18) Add $n(v\text{E} - v\text{C})$ to the weight of each constraint

Where $0 < n < 1$, $v\text{E}$ = number of violations incurred by the error, $v\text{C}$ = number of violations incurred by the correct form

For comparison's sake, the GLA can be stated as:

(19) Add n to the value of each constraint for which $v\text{E} > v\text{C}$, and subtract n from the value of each constraint for which $v\text{E} < v\text{C}$

In the GLA, the degree of difference between violations on a constraint in the error and correct form is irrelevant (as it should be for OT evaluation).

The Perceptron model has a convergence/correctness proof (Novikoff 1962), which is extended to part-of-speech tagging by Collins (2002). Collins' proof seems to further extend to an application of Perceptron with HG (Boersma and Pater 2007).

3 Noisy HG

The noisy HG account of variation (an adaptation of stochastic OT Boersma 1998; Boersma and Hayes 2001; see Boersma and Weenink 2007; Boersma and Pater 2007; Jesney 2007; Pater 2007a):

(20) i. Each time the grammar evaluates a candidate set, weighting values are perturbed by noise (sampled from a normal distribution with $SD = 2$)

ii. In repeated evaluations, this can produce variation in the choice of optima

In the Japanese loanword case, a voiced geminate *optionally* devoices in the presence of another voiced obstruent. This can be modeled with a noisy HG grammar in which the sum of the weights of the markedness constraints is sufficiently close to being equal to the weight of IDENT-VOICE:

(21) Variation produced by noisy HG

Mean weight		100	50	50	$\mathcal{H}$
	/doggu/	IDENT-VOICE	*VCE-GEM	*2-VOICE	
50%	[doggu]		-1	-1	-10 (μ)
50%	[dokku]	-1			-10 (μ)

4 Perceptron with noisy HG

Kawahara (2006) empasizes that the devoicing pattern is emergent: it has entered Japanese in recent borrowings, and there would have been no evidence in the learners' input for a difference between the forms that show devoicing, and those that do not. This pattern is in fact a predicted stage of gradual learning in HG.

We start with constraint values that are appropriate for the pattern of voicing in the native vocabulary, in which there is no voiced geminates, and only a single voiced obstruent in each word:

(22) Voicing in native Japanese phonology

Mean weight	100	50	$\mathcal{H}$
/bobu/	*2-VOICE	IDENT-VOICE	
[bobu]	-1		-100 (μ)
☞ [bopu]		-1	-50 (μ)

Mean weight	100	50	$\mathcal{H}$
/webbu/	*VCE-GEM	IDENT-VOICE	
[webbu]	-1		-100 (μ)
☞ [weppu]		-1	-50 (μ)

Praat (Boersma and Weenink 2007) implements an error-driven learner with HG evaluation, and the update rule in (18). The details of this simulation:

(23) i. Learning data: mappings /bobu/ →[bobu], /webbu/ →[webbu] and /doggu/ →[doggu] with equal frequency

ii. The initial weightings: as in (22)

iii. Learning rate (Praat's plasticity, n in (18)): 0.1.

After 1100 pieces of learning data, the weightings were as shown in (24):

(24) Emergent variable Japanese loanword devoicing

	Mean weight	107.72	53.65	$\mathcal{H}$
	/bobu/	IDENT-VOICE	*2-VOICE	
100%	[bobu]		-1	-53.65 (μ)
	[bopu]	-1		-107.72 (μ)

	Mean weight	107.72	52.50	$\mathcal{H}$
	/webbu/	IDENT-VOICE	*VCE-GEM	
100%	[webbu]		-1	-52.50 (μ)
	[weppu]	-1		-107.72 (μ)

	Mean weight	107.72	52.50	53.65	$\mathcal{H}$
	/doggu/	IDENT-VOICE	*VCE-GEM	*2-VOICE	
67.5%	[doggu]		-1	-1	-106.15 (μ)
32.5%	[dokku]	-1			-107.72 (μ)

This simulation is a gross abstraction from the actual process of loanword adaptation, and does not explain why this grammar seems to be stable in current Japanese. However, the fact that this pattern emerges as a stage in learning should make the necessary further elaboration tractable.

5 Testing gradual learning algorithms

5.1 Testing GLAs on categorical grammars

There is not (yet) a proof of that the Perceptron update rule will convergence on a correct categorical grammar when combined with noisy HG. A method for testing learning algorithms (Boersma 2007a):

(25) i. Violation marks are randomly assigned to candidates

ii. Values are randomly assigned to constraints

iii. Optima are found using an OT ranking (no noise)

iv. The learner is provided with i. and iii., and has to find constraint values, given an even distribution of the input-output mappings in a maximum of 1, 100, 000 trials (100, 000, then success check, then 1, 000, 000 more if necessary)

v. Success checked using Praat's "Fraction correct" (100, 000 randomly chosen inputs submitted to learned grammar), with noise set to zero: Criterion is 100% correct

Common settings for all simulations reported here:

(26) i. Each language has 20 constraints and 20 sets of 20 candidates

ii. Each candidate has between 0 and 5 violations of each constraint

iii. Initial constraint values of 10, and update value of 0.1 (Praat's plasticity, n in 18)

iv. In all HG simulations, the update rule is as in (18)

v. Each learner is tested on 1000 languages

Success rates:

(27) i. *Stochastic OT + standard GLA* (Boersma and Hayes 2001; Praat's 'symmetric all', 'maximal GLA' in Boersma 1998):

963/1000

ii. *Stochastic OT + demotion only GLA* (Boersma 1998):

1000/1000

iii. *HG without noise:*

1000/1000

iv. *HG with noise:*

1000/1000

v. *Positive HG (minimum value = 1) without noise:*

1000/1000

vi. *Positive HG, with noise:*

999/1000

The failure of the standard GLA with stochastic OT to find correct rankings for randomly generated languages shows that its convergence issues are general, and are not just limited to carefully engineered problems like the one in Pater (2007a).

5.2 Testing GLAs on probability matching

The success rate for learning languages with variation can be assessed by making the following change to the procedure in (25) (Boersma 2007b):

(28) i. Noise is added in the generation of the target language, which is produced using the same type of grammar as the learner's (stochastic OT and noisy HG are *not* in a subset relation)

ii. The target language is then a distribution over candidate mappings

iii. Success is checked by generating 10, 000 outputs for each input using the target grammar and the learned grammar. The Distance score is absolute difference between probabilities of occurrence, averaged over all candidates and all candidate sets.

The following results were obtained using the same settings as in (26), but with 100 rather than 1000 target languages. Praat also implements a gradual MaxEnt learner that uses the update rule in (18); this is the stochastic gradient ascent model in Jäger (2006). MaxEnt weights had initial value of 1.0, which resulted in an amount of variation in the target language comparable to the other models of grammar with initial value 10.0 and noise 2.0.

(29) i. *Stochastic OT with symmetric all:*

Mean Distance = 0.057 (*SD* 0.005)

ii. *Stochastic OT with demotion only GLA:*

Mean Distance = 0.47 (*SD* 0.06)

iii. *Noisy HG:*

Mean Distance = 0.025 (*SD* 0.007)

iv. *Positive noisy HG:*

Mean Distance = 0.025 (*SD* 0.006)

v. *MaxEnt:*

Mean Distance = 0.046 (*SD* 0.006)

It is difficult to know if the differences between the results with Mean Distance < 0.1 are meaningful, but it is clear that stochastic OT with the demotion only GLA fails to adequately match probability distributions. As Boersma (1998) predicted, demotion only fails to converge on grammars with variation.

Stochastic OT's GLA and HG's Perceptron are comparable in their complexity, in that both are extremely simple on-line learners. However, their performance is different:

(30) i. Perceptron with noisy HG succeeds in both learning categorical choices of optima and matching probability distributions over optima

ii. The standard GLA with stochastic OT fails in learning categorical choices of optima

iii. The demotion-only version of the GLA with stochastic OT fails to match probability distributions

6 Gradual learning and gradient well-formedness

A weighted constraint grammar that is learned gradually will reflect frequency differences in the relative weights of constraints. Even when these differences don't effect the outcome in input-output mappings, they can affect acceptability scores, both for the observed forms, and for ill-formed candidates (see relatedly Zuraw 2000).

A simple hypothetical case of a frequency skewing:

(31) In onset clusters, C2 is coronal 91% of the time

I submitted the following distribution of onset clusters to Praat with the standard settings for learning (except for HG evaluation/learning).

(32) Frequency of input data

klV 91%

kwV 9%

The markedness constraints began at a weighting value of 100:

Figure 1: Observed and unattested forms evaluated by a gradually learned grammar

	ranking value	*disharmony*
C2–Oral	100.000	101.886
Faith	84.971	84.881
C2–Cor	71.193	69.836
C2–Lab	53.836	53.123

klV	C2–Oral	Faith	C2–Cor	C2–Lab	
☞ klV				*	–53.123
null		*			–84.881

kwV	C2–Oral	Faith	C2–Cor	C2–Lab	
☞ kwV			*		–69.836
null		*			–84.881

knV	C2–Oral	Faith	C2–Cor	C2–Lab	
knV	*			*	–155.009
☞ null		*			–84.881

kmV	C2–Oral	Faith	C2–Cor	C2–Lab	
kmV	*		*		–171.722
☞ null		*			–84.881

(33) C2-Oral Assign a violation mark if the second member of an onset cluster is nasal

C2-Cor Assign a violation mark if the second member of an onset cluster is not coronal

C2-Lab Assign a violation mark if the second member of an onset cluster is not labial

There was just one alternative candidate for each form, a mapping to a null output, and one faithfulness constraint 'Faith' which this mapping violated. The initial value for Faith was 10. Figure 1 shows the resulting grammar, and its evaluation of the attested forms, as well as two ill-formed onsets. Both the relative Harmony of the attested forms, and of the ill-formed clusters, is affected by the weighting C2-Cor $>$ C2-Lab.

The fact that the unattested forms are differentiated by their Harmony scores highlights an important difference between this phonological approach to gradient phonotactics, and a theory that relies only on raw segmental probability (see relatedly Moreton 2002; Albright 2006; Hayes and Wilson 2006):

(34) Generalization based on phonological constraints leads to predicted asymmetries in structures of equal frequency, based on distributional differences elsewhere in the language, and on universals that

are encoded in the constraint set

In Coetzee and Pater (2007), this approach is applied to restrictions against homorganic consonants in Arabic and Muna. The correlation between the resulting Harmony scores and O/E is stronger than that between Similarity (Frisch et al., 2004) and O/E, especially in Arabic.

7 Harmony and Acceptability

One of the fundamental arguments for HG is that Harmony scores provide a means of modeling degrees of acceptability (Legendre et al. 1990, 2006a). In the standard account (see also Keller 2000; 2006), degree of acceptability = $\mathcal{H}$. The abstract example in (35) shows how multiple violations of a single constraint leads to decreased $\mathcal{H}$, and hence decreased acceptability (see Ohala and Ohala 1986 on cumulative ill-formedness in phonotactic acceptability judgments; cf. section 1 above) .

(35) Standard HG account:$\mathcal{H}$(Output-12) > $\mathcal{H}$(Output-22)

	Weight	1.5	1	$\mathcal{H}$
	Input-1	MARK	FAITH	
☞	Output-11		-1	-1
	Output-12	**-1**		**-1.5**

	Weight	1.5	1	$\mathcal{H}$
	Input-2	MARK	FAITH	
☞	Output-21		-2	-2
	Output-22	**-2**		**-3**

Boersma (2004) points out an apparently fatal flaw in this approach: because it does not relativize acceptability to the Input, in many cases ill-formed structures will receive higher scores than well-formed ones (see also Legendre et al. 2006b: 354 for a related conceptual critique of Legendre et al. 1990, 2006a).

(36) A problem for the standard HG account: $\mathcal{H}$(Output-21) > $\mathcal{H}$(Output-12)

	Weight	3	2	1	$\mathcal{H}$
	Input-1	C1	C2	C3	
	Output-11	-2			-6
☞	**Output-12**		**-2**		**-4**

	Weight	3	2	1	$\mathcal{H}$
	Input-2	C1	C2	C3	
	Output-21		**-1**		**-2**
☞	Output-22			-1	-1

A 'real-life' example using Lombardi's (1999) analysis of final devoicing (see Boersma 2004 for syntactic examples):

(37) A problem for the standard HG account: $\mathcal{H}$([pad]) > $\mathcal{H}$([bam.bam])

Weight	3	2	1	$\mathcal{H}$
/bambam/	IDENT-ONS-VCE	*VOICE	ID-VCE	
[pam.pam]	-2		-2	-8
[pam.bam]	-1	-1	-1	-6
☞ **[bam.bam]**		**-2**		**-4**

Weight	3	2	1	$\mathcal{H}$
/pad/	IDENT-ONS-VCE	*VOICE	ID-VCE	
[pad]		**-1**		**-2**
☞ [pat]			-1	-1

To make gradient ill-formedness comparative and relative, I propose that Acceptability is the difference between the Harmony of a mapping and the most harmonic alternative:

(38) Acceptability(x) = $\mathcal{H}(x)$ - $\mathcal{H}(y)$

Where x is a candidate mapping, and y is the most harmonic mapping for the same Input, and where $x \neq y$

Using this metric, grammatical forms get positive scores, and ungrammatical ones get negative scores:

(39) Acceptability([bam.bam]) = $\mathcal{H}$([bam.bam]) - $\mathcal{H}$([pam.bam]) = **2**

Acceptability([pad]) = $\mathcal{H}$([pad]) - $\mathcal{H}$([pat]) = **-1**

Cumulative ill-formedness continues to be predicted under the revised HG Acceptability metric, as the following calculation for (35) shows:

(40) Acceptability(Output-11) = $\mathcal{H}$(Output-11) - $\mathcal{H}$(Output-12) = **-0.5**

Acceptability(Output-21) = $\mathcal{H}$(Output-21) - $\mathcal{H}$(Output-22) = **-1**

None of the three following OT accounts of gradient acceptability addresses this problem, and deals with both gradient ill-formedness and well-formedness. Examples from phonotactics:

(41) Types of gradient acceptability

i. *Gradient well-formedness*: Two attested structures in a language differ in their acceptability

Hebrew roots with identical consonants in C2 and C3 (SMM) vs. heterorganic sequences (PSM) (Berent and Shimron 1997; Berent et al. 2001)

Arabic roots with homorganic consonants of different degrees of similarity (Frisch and Zawaydeh 2001)

ii. *Gradient ill-formedness*: Two unattested structures differ in their acceptability

Cumulative ill-formedness (Ohala and Ohala 1986)

Onset clusters unattested in English (Scholes 1966; Pertz and Bever 1975; Smolensky et al. 2003; Berent et al. 2006)

Coetzee's (2004) account compares performance on the constraint hierarchy for Output candidates from different Inputs (see also Everett and Berent 1998). As such, it has the same weakness as the standard HG account:

(42) A problem for cross-tableau comparison

	Ident-Ons-Vce	*Voice	Id-Vce
/bambam/→ [bam.bam]		**	
/pad/→ [pad]		*	

Two OT accounts of gradient acceptability that are Input-relative involve considerable elaboration of the theory, and work only on one side of the ill-formed/well-formed divide:

(43) a. The account of gradient well-formedness in Pater (2005) involves the postulation of lexically specific faithfulness constraints for every lexical item, and requires the calculation of the outcome of every possible indexation to determine degree of well-formedness for a nonce word. As Hayes and Wilson (2006) point out, it does not extend to gradient ill-formedness.

b. The account of gradient ill-formedness in Boersma (2004) involves the incorporation of numerical values of constraints, and the calculation of the percentage of times each candidate wins relative to the other candidates. This account does not extend to gradient well-formedness, since by definition attested (non-variable) forms win 100% of the time.

8 Conclusions

Much remains to be done in terms of further developing and evaluating HG as a theory of phonological typology, learning and gradience. However, the results we have thus far indicate that it has significant advantages over OT in modeling learning and gradience, and that it is also viable as a framework for typological study.

References

Albright, Adam. 2006. Gradient phonotactic effects: lexical? grammatical? both? neither? Albuquerque. LSA 2006.

Becker, Michael, Christopher Potts, Rajesh Bhatt, and Joe Pater. 2007. Ot-help: Java tools for optimality theory Software package, UMass Amherst.

Berent, Iris, Daniel Everett, and J. Shimron. 2001. Do phonological representations specify formal variables? Evidence from the Obligatory Contour Principle. *Cognitive Psychology* .

Berent, Iris, and J. Shimron. 1997. The representation of Hebrew words: Evidence from the Obligatory Contour Principle. *Cognition* 64:39–72.

Berent, Iris, Donca Steriade, Tracy Lennertz, and Vered Vaknin. 2006. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* .

Bhatt, Rajesh, Joe Pater, and Christopher Potts. 2007. Harmonic Grammar with Linear Programming. Ms., UMass Amherst.

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

Boersma, Paul. 2004. A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments Ms, University of Amsterdam.

Boersma, Paul. 2007a. Glasuccessrate.praat. Available from http://www.fon.hum.uva.nl/paul/gla/. Script for Praat software.

Boersma, Paul. 2007b. Glasuccessratevar.praat. Available from http://www.fon.hum.uva.nl/paul/gla/. Script for Praat software.

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

Boersma, Paul, and Joe Pater. 2007. Testing gradual learning agorithms. Ms, University of Amsterdam and UMass Amherst.

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

Coetzee, Andries. 2004. What it means to be a loser: Non-optimal candidates in Optimality Theory. Ph. D dissertation, UMass Amherst.

Coetzee, Andries, and Joe Pater. 2007. Weighted constraints and gradient phonotactics in Muna and Arabic. Ms, University of Michigan and UMass Amherst.

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

Everett, Daniel, and Iris Berent. 1998. The comparative optimality of hebrew roots: An experimental approach to violable identity constraints. Ms, University of Pittsburgh and Florida Atlantic University.

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

Frisch, Stefan, Janet Pierrehumbert, Broe, and Michael Broe. 2004. Similarity avoidance and the ocp. *Natural Language and Linguistic Theory* 22:179–228.

Frisch, Stefan, and Bushra Zawaydeh. 2001. The psychological reality of ocp-place in arabic. *Language* 77:91–206.

Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the stockholm workshop on variation within optimality theory*, ed. Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120. Stockholm: Stockholm University.

Hayes, Bruce, Bruce Tesar, and Kie Zuraw. 2003. OTSoft 2.1. URL `http://www.linguistics.ucla.edu/people/hayes/otsoft/`, Software package developed at UCLA.

Hayes, Bruce, and Colin Wilson. 2006. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* To appear.

Jäger, Gerhard. 2006. Maximum entropy models and Stochastic Optimality Theory. In *Architectures, rules, and preferences: A festschrift for Joan Bresnan*, ed. Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, and Annie Zaenen. Stanford, CA: CSLI. To appear.

Jesney, Karen. 2007. The locus of variation in weighted constraint grammars. In *Poster presented at the Stanford Workshop on Variation, Gradience and Frequency in Phonology*. Stanford University, Stanford CA.

Johnson, Mark. 2002. Optimality-theoretic lexical functional grammar. In *The lexical basis of syntactic processing: Formal, computational and experimental issues*, ed. Suzanne Stevenson and Paola Merlo, 59–73. John Benjamins.

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

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

Keller, Frank. 2000. Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Doctoral Dissertation, University of Edinburgh.

Keller, Frank. 2006. Linear optimality theory as a model of gradience in grammar. In *Gradience in grammar: Generative perspectives*, ed. Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky. Oxford: Oxford University Press.

Legendre, Geraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An Application. In *Proceedings of the twelfth annual conference of the cognitive science society*, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 2006a. The interaction of syntax and semantics: A Harmonic Grammar account of split intransitivity. In Smolensky and Legendre (2006), 417–452.

Legendre, Géraldine, Antonella Sorace, and Paul Smolensky. 2006b. The Optimality Theory–Harmonic Grammar connection. In Smolensky and Legendre (2006), 339–402.

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

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

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

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

Moreton, Elliott. 2002. Structural constraints in the perception of english stop-sonorant clusters. *Cognition* 84:55–71.

Nishimura, Kohei. 2003. Lyman's law in loanwords. Ms, University of Tokyo.

Novikoff, A.B.J. 1962. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, 12, 615–622. Polytechnic Institute of Brooklyn.

Ohala, John J., and Manjari Ohala. 1986. Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In *Experimental phonology*, ed. John J. Ohala and Jeri J. Jaeger, 239–252. Orlando: Academic Press.

Pater, Joe. 2005. Learning a stratified grammar. In *Proceedings of the 29th boston university conference on language development*, ed. Alejna Brugos, Manuella R. Clark-Cotton, and Seungwan Ha, 482–492. Somerville, MA: Cascadilla.

Pater, Joe. 2006. Additive optimization and phonological typology. URL `http://people.umass.edu/pater/AO+typology.pdf`, Handout from a talk presented at the MIT/UMass Phonology Workshop, February 11.

Pater, Joe. 2007a. Gradual learning and convergence. *Linguistic Inquiry* To appear.

Pater, Joe. 2007b. Review of Smolensky and Legendre 2006. *Phonology* To appear.

Pater, Joe, Rajesh Bhatt, and Christopher Potts. 2007. Linguistic optimization. Ms, University of Massachusetts, Amherst.

Pertz, D. L., and T. G. Bever. 1975. Sensitivity to phonological universals in children and adolescents. *Language* 51:149?162.

Potts, Christopher, Michael Becker, Rajesh Bhatt, and Joe Pater. 2007. HaLP: Harmonic grammar with linear programming, version 2. Software available online at http://web.linguist.umass.edu/~halp/.

Prince, Alan. 2002. Anything goes. In *New century of phonology and phonological theory*, ed. Takeru Honma, Masao Okazaki, Toshiyuki Tabata, and Shin ichi Tanaka, 66–90. Tokyo: Kaitakusha. ROA-536.

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

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

Scholes, Robert. 1966. *Phonotactic grammaticality*. The Hague: Mouton.

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

Smolensky, Paul, Lisa Davidson, and Peter Jusczyk. 2003. The Initial and Final States: Theoretical Implications and Experimental Explorations of Richness of the Base. In *Fixing priorities: constraints in phonological acquisition*, ed. Rene Kager, Joe Pater, and Wim Zonneveld. Cambridge: Cambridge University Press.

Smolensky, Paul, and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to Optimality-Theoretic grammar*. Cambridge, MA: MIT Press.

Soderstrom, Melanie, Donald W. Mathis, and Paul Smolensky. 2006. Abstract genomic encoding of universal grammar in optimality theory. In *The harmonic mind: From neural computation to optimality-theoretic grammar. volume 2: Linguistic and philosophical implications.*, ed. Paul Smolensky and GÈraldine Legendre, 403–471. MIT Press.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.

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

Zuraw, Kie. 2000. Patterned exceptions in phonology. Doctoral Dissertation, UCLA.