Varieties of Noisy Harmonic Grammar

Bruce Hayes
University of California, Los Angeles

1. Background: Stochastic constraint-based grammar frameworks in modern linguistics

The key innovation of Optimality Theory (Prince and Smolensky 1993) was its GEN-plus-EVAL architecture: GEN enumerates candidates, and EVAL, consisting of a set of constraints, selects the winning candidate from GEN as the output. This conception naturally led to the idea of a stochastic version of the theory, in which EVAL outputs not one single candidate but rather a probability distribution over GEN. Such a framework would provide a natural account of free variation and related phenomena: we get multiple outputs when the conflict between constraints is not completely resolved. Early such frameworks included the Floating Constraint model (Nagy and Reynolds 1997), Partial Ordering OT (Anttila 1997) and Stochastic OT (Boersma 1997), and with time still other approaches were put forward.

The pursuit of such frameworks is empirically well-motivated because gradience in language is so widespread (see, e.g. Bod et al. 2003, Fanselow et al. 2006). Everywhere we look, if we look carefully, we find free variation in the outputs of grammar and gradience in well-formedness intuitions. The work of Zuraw (2000, 2010) opened up a new domain of gradience, quantitative phonological patterns in the lexicon, often perceived accurately by language learners, with the ambient probability distributions accurately reproduced by them under experimental probing. Empirical work continues to support this "Law of Frequency Matching" (Hayes et al. 2009) as a baseline prediction for human phonological learning. All of these phenomena need an appropriate framework for formal analysis. [1]

The earlier phases of research on stochastic grammar frameworks were tinged with a practical orientation: it was felt to be important simply to establish that the frameworks really could work in nontrivial cases, e.g. by Boersma and Hayes (2001:46). It was also felt to be important to assess the learning algorithms offered in tandem with new stochastic frameworks. Such research is exemplified by the counterexample Pater (2008) discovered to the Gradual Learning Algorithm for Stochastic OT; or Goldwater and Johnson's (2003) pointing out the sensibleness of using maxent based on its strong and reliable mathematical foundations. Soon, however, purely scientific goals entered the debate: different stochastic frameworks treat the same gradient data in different ways and make different empirical predictions, so that the choice of framework is really part of linguistic theory. Early on, Jesney (2007) demonstrated the substantial differences between Noisy Harmonic Grammar and maxent grammars in their treatment of consonant deletion for a CVCC input. Similarly, Boersma and Pater (2008/2016:410-412) compared versions of Harmonic Grammar in which constraint weights are either allowed or not allowed to go below zero. This article is intended as a contribution to this line of research: I put forth some cases in which different varieties of Noisy Harmonic Grammar, as well as maximum entropy grammars, make different predictions.

* Thanks to Kie Zuraw for substantial advice and assistance, and to Adam Albright, Joe Pater, Brian Smith, and talk audiences at the UCLA Phonology Seminar and AMP for their helpful feedback.

[1] My sense is that there remains some inertia among linguists in the recognition of variation as an empirical reality. I conjecture that this may be the consequence of traditional research methodologies of the field. It is still often the case that fieldworker and theorist never meet, and the latter takes the former's first-pass best estimate as the analytical target. Different results are often obtained if the theorist has access to corpora, or performs experiments. Another factor may be the widespread use of problem sets (e.g. in first-year graduate training) in which the data have been intentionally cleansed of variation.

Proceedings of *AMP 2016*

2. Background on the frameworks examined

2.1 *Noisy Harmonic Grammar* Noisy Harmonic Grammar (hereafter "NHG") is stochasticized Harmonic Grammar. Harmonic Grammar, which has an ancestry that predates OT (Legendre et al. 1990, Legendre et al. 2006, Pater 2008/2016, Potts et al. 2010), uses the same GEN-cum-EVAL architecture, but instead of ranking the constraints, it assigns them numerical weights reflecting their relative strength. In the nonstochastic version of the framework, winning candidates are selected as follows. For each candidate's row in the tableau, one multiplies violation counts by corresponding constraint weights and add up the total across constraints. This yields the **harmony**, a kind of penalty score.[2] The winning candidate is the least penalized one; i.e. the one with the lowest harmony. In (1), we first see an OT-like tableau, with two constraints and two candidates, along with the weights of the constraints. Tableau (1b) shows how the basic computations are carried out, and it can be seen that Candidate 1, with the lesser harmony penalty, emerges as the winner.

(1) *Illustration of nonstochastic Harmonic Grammar*

a. /Input/	CONSTRAINT1	CONSTRAINT 2
weights:	2	1
Candidate 1		**
Candidate 2	*	*

b. /Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weights:	2	1	
☞ Candidate 1		** × 1 = 2	0 + 2 = **2**
Candidate 2	* × 2 = 2	* × 1 = 1	2 + 1 = **3**

In what follows I will discuss stochastic versions of this theory.

2.2 *Classical Noisy Harmonic Grammar* In what I will call Classical NHG, put forth by Boersma and Pater (2008/2016), Harmonic Grammar is stochasticized as follows. At each "evaluation time" (moment of application of the grammar), every constraint weight is separately perturbed upward or downward by a random amount drawn from a Gaussian distribution with mean 0 and a uniform standard deviation. This perturbation factor is called the **noise**; I will here employ a noise value of 1. Because of noise, on different evaluation times, there can often can be different winners, depending on what random noise values happen to have been assigned.

For example, in (2) noise values (designated here as N_n) are added to each weight, influencing the computation of harmony. Candidate 2 is penalized by a greater "base value" of harmony (3 vs. 2), making it mostly likely that Candidate 1 will win. But if the noise values happen to be such that N_2 exceeds N_1 by more than one, then Candidate 2 will win instead. The math when worked out shows that Candidate 2 will in fact win 24.0% of the time. The frequency pattern is shown in (2) informally with pointing fingers of different sizes.

[2] Different scholars use different versions of Harmony, varying only in a negative sign (for some Harmony is a negative quantity). The nomenclature/choice of signs followed here is taken from Wilson (2006).

(2) *Illustration of Classical Noisy Harmonic Grammar*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	$2 + N_1$	$1 + N_2$	
☞ Candidate 1		$** \times (1 + N_2)$	$2 + 2N_2$
☞ Candidate 2	$* \times (2 + N_1)$	$* \times (1 + N_2)$	$3 + N_1 + N_2$

The probability of 24.0% can be calculated, for instance, by performing various integrations over Gaussian distributions, a procedure that becomes laborious for all but simple cases (see Zuraw 2000:105-113). A convenient alternative is to use the Monte Carlo method: one runs the same grammar, say, 100,000 times, counting up winners, and the values obtained, divided by 100,000, serve as a reasonable estimate of the probabilities generated by the grammar.

The procedure established in Classical Noisy Harmonic Grammar represents just one choice from a whole set of logical possibilities for introducing noise into Harmonic Grammar. I conjecture that this choice was carried over from the pioneering first theory that used noise to stochasticize a constraint-based framework, namely Boersma's (1997) Stochastic OT. Just as Stochastic OT adds noise to the "ranking values" of constraints and then uses the result to pick a winner by the rules of nonstochastic OT, so Classical NHG adds noise directly to the constraint weights and picks a winner by the rules of nonstochastic Harmonic Grammar.

The alternatives to Classical NHG that I will consider here derive from just where in the system one "adds in the noise". I will explore variants using cell-specific noise (§2.3) as well as variants that add the noise in late, e.g. after the multiplication of weights by violation counts (§3.3).

2.3 *NHG with cell-specific noise* Goldrick and Daland (2009) suggest that we can profitably explore more fine-grained assignments of noise than in Classical NHG. Their proposal is actually made in a connectionist implementation of Harmonic Grammar that has other implications as well (for instance, different weights for $+ \rightarrow -$ Faithfulness mappings than $- \rightarrow +$), so I will here follow what is in context a more modest change in Classical NHG. Rather than perturbing the original constraint weights, we instead install a fresh noise value for every cell. For now, we assume that cells with no violations are not given a noise value, though this assumption will be revised later on. Here is an example:

(3) *Illustration of Noisy Harmonic Grammar with cell-specific noise*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	2	1	
☞ Candidate 1		$* \times (1 + N_2)$	$2 + 2N_2$
☞ Candidate 2	$* \times (2 + N_1)$	$* \times (1 + N_3)$	$3 + N_1 + N_3$

Let us compare the harmony values from the tableaux of (2) and (3): the Classical theory assigns Candidate 2 a perturbed harmony of $3 + N_1 + N_2$ whereas the theory with cell-specific noise assigns it a perturbed harmony of $3 + N_1 + N_3$; for both theories the harmony of Candidate 1 is $2 + 2N_2$. We can think of (3) as displaying "cell-granularity" and (2) "constraint-granularity". This makes differing predictions, as will be illustrated shortly.

2.4 *Maxent grammars* Before proceeding to this, I add one more theory to the list, namely **maximum entropy Harmonic Grammar** (Smolensky 1986, Goldwater and Johnson 2003, Wilson 2006, Jaeger 2007, Hayes and Wilson 2008); "maxent" for short. Maxent, like NHG, carries forward the GEN-cum-EVAL architecture of Optimality Theory, and also employs Harmony as the basis of the computations. However, in maxent, there is no noise factor. Instead, a simple procedure directly computes the output probabilities from the harmonies of the candidates. The three steps of this procedure are given in (4), shown in the computation of the probability of a hypothetical candidate *x*. From Harmony we first compute "eHarmony"

by negation and exponentiation;[3] by summing the eHarmonies of all candidates we compute the normalization term Z, and the probability of *x* is specified as the share of its eHarmony in Z.

(4) *Computing probability from harmony in maxent*

a. eHarmony: $e^{-\text{Harmony}(x)}$ (*e* to the minus Harmony of candidate *x*)

b. *Z*: Σ_j eHarmony(j) (sum eHarmony over all candidates)

c. probability: $\frac{\text{eHarmony}(x)}{Z}$ (candidate *x*'s share of Z)

These computations result in what I believe to be intuitively sensible behavior in maxent grammars. First, candidates penalized by greater harmony receive lower probability. Second, it follows from how harmony is computed that constraints with larger weights will have a greater role in lowering probability of the candidates that violate them, and that multiple violations will have more effect in lowering the probability of the violating candidate. Third, the exponentiation involved in calculating eHarmony has an important consequence, namely that altering harmony values has different effects on different regions of the probability scale: it requires a large shift in the harmony penalty to alter probabilities that are already close to zero or one, but only small differences in harmony are needed make an appreciable probability shift in the medial regions near 50%. This effect will be seen below (§3.3) when we study the sigmoid curves that maxent generates.

For the tableau of (1), maxent generates probabilities as shown in (5).

(5) *Illustration of maxent*

/Input/	CONST1	CONST2	Harmony	eHarmony	Z	Probability
weights:	2	1			.135 + .05 = .185	
Candidate 1		**	2	e^{-2} = .135		.135/.185 = **.73**
Candidate 2	*	*	3	e^{-3} = .05		.05/.185 = **.27**

3. Examining the behavior of the various frameworks

We now have three stochastic frameworks in hand (two versions of Noisy Harmonic Grmamar plus maxent); more will be added in what follows. The goal of this main section is to confront these various frameworks with simple, representative cases in which they behave differently. The goal is to provide schematic cases that have at least some hope of being fulfilled by real-life empirical examples, bringing the theories to the test. My examples will involve harmonic bounding, the local-optionality problem, and the competition of gradient against categorical constraints.

3.1 *Harmonic bounding* This section reviews results already presented in Prince (2002), Jesney (2007), Goldrick and Daland (2009), and Boersma and Pater (2016). A very simple case of harmonic bounding is given in tableau (6).

[3] I borrow a jocular, but clear and useful, term from pedagogical materials created by Colin Wilson.

(6) *Harmonic bounding*

a. *Simple violations*

/Input/	CONSTRAINT1	CONSTRAINT 2
Candidate 1		*
Candidate 2	*	*

b. *Harmony computation in Classical NHG*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	W_1	W_2	
☞ Candidate 1		$* \times (W_2 + N_2)$	$(W_2 + N_2)$
Candidate 2	$* \times (W_1 + N_1)$	$* \times (W_2 + N_2)$	$(W_1 + N_1) + (W_2 + N_2)$

Candidate 2 has a proper superset of the violations of Candidate 1, and by the now-standard nomenclature of OT we say that Candidate 1 **harmonically bounds** Candidate 2.

Under a suitable interpretation, Classical NHG preserves the OT principle that harmonically-bounded candidates never win; i.e. they are assigned the probability zero. The essential auxiliary assumption is that weights, even after perturbation under noise, are always kept positive (Boersma and Pater 2016:401-412), a matter that can be imposed by fiat. To make clear what weights are being kept positive by fiat, I will enclose expressions like $(W_n + N_m)$, showing the perturbed value, with parentheses in the tableaux. In (6), we see that Candidate 2 always has a greater harmony penalty than Candidate 1, since the difference in harmony is $(W_1 + N_1)$, which as a perturbed harmony value must be greater than zero. More generally, no harmonically bounded candidate can ever win in Classical NHG when the no-negative weight proviso is in effect.

But in cell-granular NHG, as Goldrick and Daland pointed out, harmonically-bounded candidates do not always necessarily lose, even under a no-negative-weight regimen. For our schematic case, the calculations are as in (7):

(7) *Lack of harmonic bounding in cell-granular HNG*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	W_1	W_2	
☞ Candidate 1		$* \times (W_2 + N_1)$	$(W_2 + N_1)$
☞ Candidate 2	$* \times (W_1 + N_2)$	$* \times (W_2 + N_3)$	$(W_1 + N_2) + (W_2 + N_3)$

Since there are separate noise values for each cell, it is quite possible that these will be such that $(W_2 + N_1) > (W_1 + N_2) + (W_2 + N_3)$, even if all weights are required to be positive, so it is possible for Candidate 2 to win. The fact that Candidate 2 bears the additional penalty of W_1 not borne by Candidate 1 means that Candidate 1 will be *more likely* to win. In general, cell-granular NHG requires that harmonically bounded candidates get a lower probability than the rival candidates that bound them, but they do not get shut out entirely as they do under the all-weights-positive version of Classical NHG. This can be thought of as the "relative" version of harmonic bounding, and is characteristic of theories that do not impose absolute harmonic bounding.

Maxent also has only the relative version of harmonic bounding. As Jesney (2007) pointed out, this can be deduced directly from the maxent formulae, given above in (4). The key point is that *any candidate whatsoever* must get at least some positive probability, which of course totally excludes harmonic bounding. The reason is this: *e* to any number is positive, and so eHarmony is always positive. Since Z is a sum of eHarmonies, it too is positive, and therefore the probability assigned to any candidate is positive. The always-positive-probability is not as drastic a move as it might seem, because in an adequate analysis

the truly bad candidates are assigned infinitesimally small probabilities; few sensible scholars, I think, will worry about bad candidates whose positive probability is 10^{-50}. For why maxent observes "relative" harmonic bounding, note that if Candidate 1 harmonically bounds Candidate 2, then Candidate 2 will have the higher Harmony penalty, and hence a lower eHarmony (negation plus exponentiation), and hence a lower probability.[4]

Summing up so far: Classical HNG, suitably constrained to keep weights positive, is the only framework so far that respects classical (absolute) harmonic bounding. Cell-granularity NHG and maxent respect only a relative version (the bounded candidate gets lower probability than its bounder).

3.2 *A special case of harmonic bounding: Local optionality* Local optionality is a long-standing challenge for OT and other constraint-based theories, and has accumulated a modest research literature including Vaux (2003, 2008), Riggle and Wilson (2005), Kaplan (2011, 2016), and Kimper (2011). Intuitively, the idea is that the same phonological process, guided by the same constraints, is applicable at more than one place in the input. For example, if we had a language that optionally voiced intervocalic /p/ to [b], then an underlying form /apapa/ might yield four plausible candidates: [apapa], [apaba], [abapa], and [ababa]. The middle two of these, with internally mismatched outcomes, are the cases that illustrate local optionality; the other two illustrate application or non-application of voicing "in lockstep."[5]

The literature just cited describes a small set of useful empirical cases of local optionality, including French schwa deletion, Makonde vowel harmony, Bengali intonation phrasing, and Pima reduplication. To my knowledge, in none of these cases is it (yet) feasible to test the differences I am about to demonstrate. Indeed, for the case of French schwa deletion, the existence of "extraneous" constraints (tied to particular morphemes; Kaplan 2016) suggest that a pure test of the differences I will illustrate may be outright impossible. Nevertheless, the differences are in principle substantial and it seems useful to keep them in mind as further empirical cases arise.

For clarity, let us scale up our schematic /apapa/ input to /apapapapa/, with four /p/'s, and group the candidates into classes according to how many /p/ become [b]. Adopting standard Markedness and Faithfulness constraints, we get a tableau like (8).

(8) *Local optionality in /apapapapa/*

/apapapapa/	IDENT(voice)	*VpV
[apapapapa]		****
[apapapa**b**a] et al.	*	***
[apapa**b**a**b**a] et al.	**	**
[apa**b**a**b**a**b**a] et al.	***	*
[a**b**a**b**a**b**a**b**a]	****	

The tableau shows double pyramids of asterisks. These are characteristic in multiple application cases, since there is a direct tradeoff of Markedness and Faithfulness violations at every application locus.

Classical NHG (assuming the no-negative weights proviso) can derive only [apapapapa] and [abababababa], the lockstep candidates. At any evaluation time either IDENT(voice) or *VpV will have the lower perturbed weight, and the candidate that concentrates its violations on that constraint will emerge as the winner. As previous authors have pointed out, this is harmonic bounding, or more specifically the "collective" harmonic bounding described by Samek-Lodovici and Prince (1999); the lockstep candidates jointly, rather than singly, exclude all the others.

In contrast, maxent and cell-granularity NHG, which do not respect harmonic bounding, *can* assign probability to the medial candidates, as we will see. But I would like to take the question one step further:

[4] A minor correction: if the constraints on which Candidate 1 and Candidate 2 differ all have zero weights, then of course they will have identical Harmonies and identical probabilities. Candidate 2 can never have a *higher* probability than Candidate 1.

[5] A real-life counterpart of this language, Warao (Osborn 1966), is claimed to have obligatory across-the-board-or-none application; this may reflect the effect of style, discussed below in §4.3 below.

what sort of *probability distributions* do the two theories characteristically assign? To see this, let us scale up the schematic input form to six intervocalic /p/, i.e. underlying /apapapapapapa/. I will assume for concreteness that the process of intervocalic voicing applies with a probability of 50%, and that there is no dependency across loci. Tableau (9) is a sketch summary of the 64 equiprobable candidates.

(9) *A tableau with six /p/ and 64 candidates*

Input	*Candidate*	*Freq.*	*VpV	IDENT(voice)
/apapapapapapa/	apapapapapapa	1	******	
	apapapapapaba	1	*****	*
	apapapapabapa	1	*****	*
	apapapabapapa	1	*****	*
	apapabapapapa	1	*****	*
	apabapapapapa	1	*****	*
	abapapapapapa	1	*****	*
	apapapapababa	1	****	**
	apapapabapaba	1	****	**
	(48 not listed)	…	…	…
	abababababapa	1	*	*****
	abababababapa	1	*	*****
	ababababababa	1		******

For the case of maxent, I submitted a file embodying (9) to the Maxent Grammar Tool (Wilson and George (2009). Under suitable conditions,[6] we get the most minimal analysis conceivable: both constraints get weighted at zero, and under this condition the data are in fact predicted perfectly — equiprobability across all 64 candidates. This is, indeed, why the theory is called maximum entropy; it makes minimal commitments (maximum randomness) in the absence of any sort of pattern in the training data.

Let us turn now to cell-granular NHG. Under our assumption that neither [p] nor [b] is favored overall, we can simplify the analysis, since this means that W_{*VpV} and $W_{ID(voice)}$ must be equal; call this number W. Moreover, every candidate will have a basic (non-perturbed) harmony penalty of 6W (see tableau (9)), which will cancel out in the harmony comparisons. This means is it safe to assume that W is fairly big (say, about 5), which has the beneficial effect of keeping perturbed weights positive essentially all the time without any intervention on our part. Now we can offer a simple picture of the relative harmonies (with W left out) for some representative candidates:

(10) *Sources of harmony for the candidate types (W omitted)*

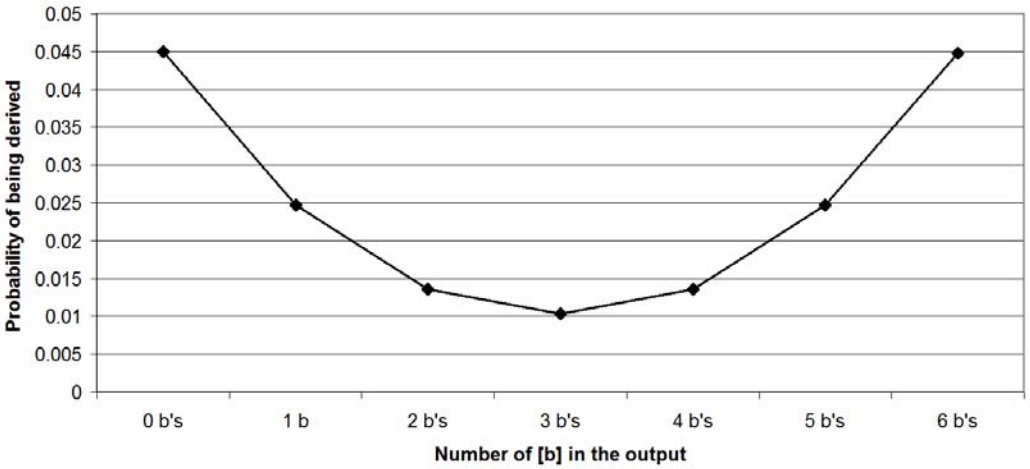
Candidate	*VpV	IDENT(voice)
apapapapapapa	$6N_1$	
apapapapapaba, other candidates with 1 [b][7]	$5N_2$	N_3
apapapapababa, other candidates with 2 [b]	$4N_4$	$2N_5$
apapapabababa, other candidates with 3 [b]	$3N_6$	$3N_7$
apapababababa, other candidates with 4 [b]	$2N_8$	$4N_9$
apabababababa, other candidates with 5 [b]	$1N_{10}$	$5N_{11}$
ababababababa		$6N_{12}$

[6] Specifically, a general penalty for constraint weights (Gaussian prior), often employed for maxent, is in effect.

[7] Bear in mind that every candidate gets its own noise, so really there are 2 x 64 − 2 = 126 noise values, of which only 12 are shown.

From this it is not hard to compute the predicted frequency of each candidate type (I used my "OTSoft" software[8] to do this). The result is given in the graph (11), which portrays a U-shaped distribution: some probability is allocated to all candidates, but with priority to the extremes, and lowest probability to the midpoint. It should be borne in mind that the probabilities given in the graph are not the aggregate probabilities assigned to (say) all three-[b] candidates, but to individual members of the 64-candidate set.

(11) *U-shaped distribution derived by cell-granularity NHG*



For reference, I propose a name for the phenomenon seen here: **upsilonism**, which is defined as a situation in which a system must generate a U-shaped probability distribution as a consequence of its architecture. It should be clear that for cell-granularity NHG, upsilonism is an absolute limitation: even if we used an algorithm to train it on a flat distribution, it could only learn a U-shaped one; given the single parameter W, it lacks the freedom to do otherwise.

Upsilonism was first discovered by Jesney (2007) in her study of CCVC syllable typology; there too, Classical NHG creates an upsilonic distribution, with the maximally faithful (CCVC) and unfaithful (CV) candidates at the extremes receiving higher probabilities. Our six-[p] simulation shows the same pattern in more extravagant form.

The presence of upsilonism in Classical NHG can be explained intuitively on the basis of chart (10). If, at some evaluation time, the noise factor N_1 happens to come out low, then the [apapapapapapa] will be likely to win, for it benefits sixfold from this lowness. At the other end of the scale, [abababababa] benefits sixfold whenever N_{12} comes out low. But the medial candidates will benefit less (fivefold, fourfold, etc.) from their lucky moments. They will have trouble standing out, the more so the closer to the center they lie.

Maxent is also biased, but in the opposite direction: all 64 candidates have the Harmony value 6W and this will generate a flat distribution no matter what choice is made for W; the distribution of the training data is irrelevant. Thus we might say that maxent shows a bias for **rectilinearism**.

I put forth this case, following Jesney, as a pattern that might someday lead to empirical testability: we need multiple-locus phonology with opposed constraints of Markedness and Faithfulness and a fairly symmetrical overall distribution, and essential non-interference from other constraints or aspects of speaking style or rate (this is perhaps a tall order). In such cases, irrespective of the overall training frequencies, classical NHG allocates all probability to the extremes, cell-granular NHG demands upsilonism, and maxent requires rectilinearism.

3.3 *Post-multiplicative vs. pre-multiplicative noise addition* Let us turn now to another special property of Classical NHG: because the noise is added in at the very beginning, it gets *multiplied by violation count* during the calculation of the perturbed harmony. Tableau (12) illustrates this; see in particular the terms $3N_2$, $2N_1$, and $2N_2$.

[8] Version 2.5: www.linguistics.ucla.edu/people/hayes/otsoft/

(12) *Multiplication of noise in Classical NHG*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	$2 + N_1$	$1 + N_2$	
☞ Candidate 1	$* \times (2 + N_1)$	$*** \times (1 + N_2)$	$5 + N_1 + 3N_2$
☞ Candidate 2	$** \times (2 + N_1)$	$** \times (1 + N_2)$	$6 + 2N_1 + 2N_2$

This is of course not a conceptual necessity. We could just as easily compute and store a noise value for each constraint, and add in this noise *after* weights have been multiplied by violations. We could call this system "Noisy HG with post-multiplicative noise"; it is illustrated in (13).

(13) *NHG with post-multiplicative noise*

/Input/	CONSTRAINT1	CONSTRAINT 2	Harmony
weight:	2	1	
☞ Candidate 1	$(* \times 2) + N_1$	$(*** \times 1) + N_2$	$5 + N_1 + N_2$
☞ Candidate 2	$(* \times 2) + N_1$	$(** \times 1) + N_2$	$6 + N_1 + N_2$

For the moment, let us adopt what is arguably the simplest implementation of this system: the noise is added in even when the number of violations is zero, and we will not bother with any provision to keep weights above zero. We will consider more complex implementations in §3.4 below.[9]

Let us now ponder a scenario in which the difference between pre-multiplicative and post-multiplicative noise will produce an observable difference in model predictions. I will assume a set of parallel binary competitions in which two viable candidate types, 1 and 2, compete for each input. Constraint 1 will always penalize Candidate 1 and will always incur just one violation. Constraint 2 penalizes Candidate 2 and is a **scalar** constraint, with a range of integer values, penalizing Candidate 2 according to how many of some property it has.

This scenario has real-world analogues pursued in earlier research (Hsu and Jesney 2016, McPherson and Hayes 2016). In McPherson and Hayes's study of Tommo So vowel harmony, the phenomenon is vowel harmony; the single-violation constraint is IDENT(feature) for each harmonizing feature, and the scalar constraint is AGREE(feature) for each of seven degrees of "morphological closeness," formalizing the morphological levels of classical Lexical Phonology (Kiparsky 1982); inner levels create more violations of AGREE.

For purposes of illustration I will set the number of values for the scalar constraint to 20; this is probably too high a number to be matched in any sort of empirical phenomenon but will illustrate the qualitative form of the predictions most clearly. The violation tableau is given in (14).

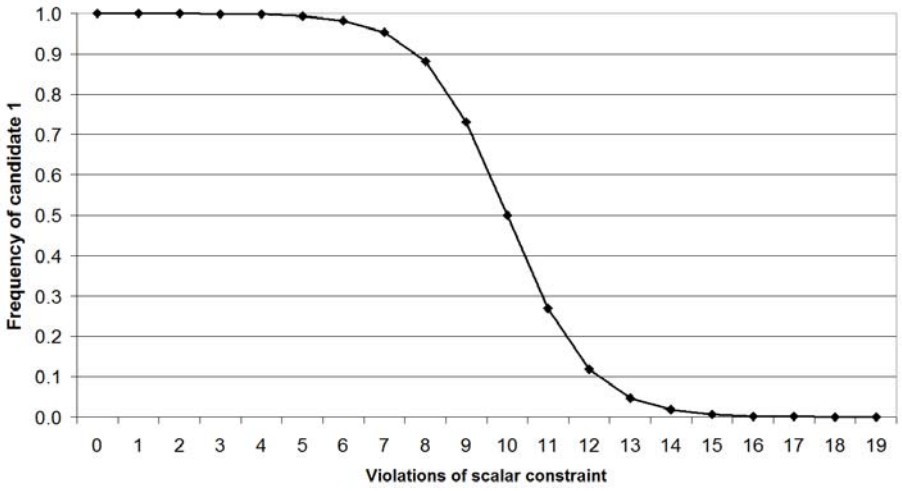
[9] Tableau (13) can be reconceptualized quite differently. Observe that every candidate is penalized by a value that consists of its basic harmony, plus the sum of one noise factor for each constraint (in (13), $N_1 + N_2$). The sum of a set of Gaussian distributions is itself a Gaussian. So (13) is very little different from a system of **candidate noise**, in which we first compute non-stochastic Harmony, then add a single noise factor to every candidate. The only difference is that in (13) the overall noise level depends on the number of constraints in the grammar. Whatever properties hold of post-violation, noise-for-all-cells, negative-weights-ok NHG will usually hold as well for candidate-noise NHG. I find candidate-noise NHG conceptually appealing because it dispenses entirely with the complexities associated with negative constraint weights; the weights are not perturbed at all. Thanks to Adam Albright and Brian Smith for pointing out candidate noise to me as a possibility.

(14) *Schematic example: scalar constraint pitted against constant constraint*

		SCALAR CONSTRAINT	CONSTANT CONSTRAINT
Input 1	Candidate 1	0	
	Candidate 2		1
Input 2	Candidate 1	1	
	Candidate 2		1
Input 3	Candidate 1	2	
	Candidate 2		1
Input 4	Candidate 1	3	
	Candidate 2		1
…	…	…	…
Input 20	Candidate 1	19	
	Candidate 2		1

Calculating with spreadsheets and OTSoft, I examined how the probability of Candidate 1 goes down as we increase the number of violations of SCALAR CONSTRAINT. We consider maxent first; chart (15) gives the result when $W_{scalar} = 1$, $W_{constant} = 10$.

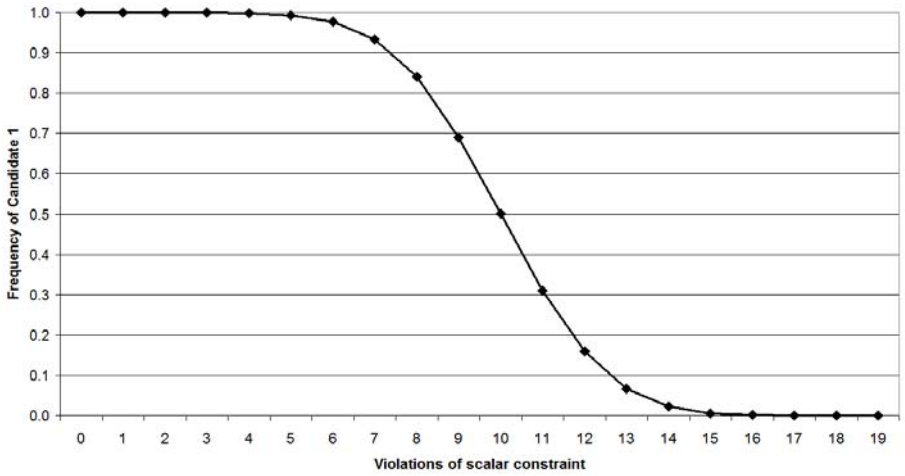
(15) *Scalar against constant constraint: curve derived under maxent*



This is a **sigmoid** curve, and its mathematical properties are well understood. As a version of the logistic function, it is symmetrical about the point of 50% probability; it asymptotes at one in the negative direction, and it asymptotes at zero in the positive direction.

Turning next to NHG with post-multiplicative noise, we can see the behavior of this theory in chart (16); it assumes $W_{Scalar} = 2$, $W_{Constant} = 20$ (twice the maxent weights).

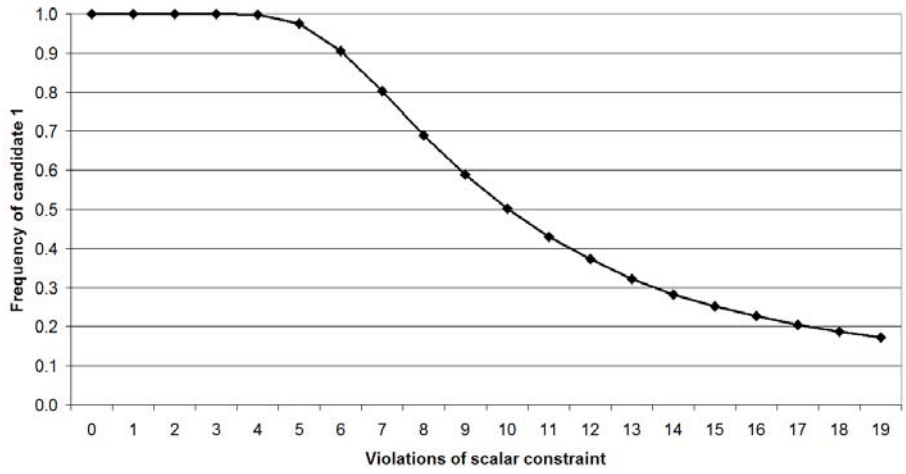
(16) *Scalar against constant constraint: sigmoid curve derived under NHG with post-multiplicative noise*



Here again we have a sigmoid, but with a different mathematical basis; it is the cumulative distribution function for the normal distribution. Its shape is, however, amazingly similar to the logistic function: it too is symmetrical about the 50% point and asymptotes at one and zero.

In Classical NHG, with pre-multiplicative noise, the result obtained is strikingly different. In (17), the weights are again 2 for the constant constraint and 20 for the scalar constraint.

(17) *Scalar against constant constraint: sigmoid curve derived under Classical NHG*



Here, the curve is asymmetrical, and although it asymptotes at one on the left, it does not asymptote at zero on the right (the rightward asymptote turns out to be about 0.07). The origin of this asymmetry lies in the broadening of the noise distribution that results when noise gets multiplied — the broadened distribution means that the range of perturbed weights for the scalar constraint continues to overlap substantially with the range of constant weights no matter how high the violation count goes, so that the right tail of the curve can never reach zero.

Summing up, I believe is it the case that all versions of NHG that multiply noise by violation count derive asymmetrical curves that fail to asymptote to zero as violations increase. Versions of NHG that do not multiply noise by violation count derive symmetrical curves asymptoting at zero and one, as does maxent.

3.4 *Revisiting the three simulations under a broader range of assumptions* The previous sections introduced or reviewed four free "parameters" for NHG: whether noise is constraint-granular or cell-

granular, whether noise is pre- or post-multiplicative, whether empty cells get noise, and whether weights can go negative. This yields a substantial group of variant NHG frameworks, which I have checked out against all three schematic cases examined here; see Supplemental Materials.[10] For harmonic bounding (§3.1), things seem fairly simple: it holds when noise is both constraint-granular and non-negative. For scalar vs. constant constraints (§3.3), the choice of symmetrical/zero-asymptote vs. asymmetrical/nonzero asymptote sigmoids is likewise straightforward, depending entirely on whether or not noise addition is post-multiplicative.

The local-optionality problem (§3.2) proves more complicated, and the outcomes observed so far (harmonic bounding, upsilonism, and rectilinearism) are not the only possibilities. We will also see "earlike" patterns, which are rectilinear except for upward or downward deviations at the extremes, as shown in (18).

(18) *Probability distributions with "ears"*

a. *Downward* b. *Upward*

The full set of patterns is as follows.

- First, if we continue to hold to the assumptions made §3.3 — noise added even for zero-violation cells, and no ban on negative weights — then upsilonism emerges as the consequence of pre-multiplicative noise, and rectilinearism results from post-multiplicative noise. This makes sense in light of the discussion in §3.2 of how pre-multiplicative noise creates a special benefit for candidates that are at the periphery of the scale.
- If violation-free cells get no noise, special effects are observed. The crucial empty cells can be seen in (9); they are found only for the "peripheral" candidates, i.e. at *VpV in the fully-faithful candidate [apapapapapapa] and at IDENT(voice) for the fully-unmarked [abababababaa].
  - In cell-granularity NHG, the peripheral candidates will have one Noise factor instead of two, and thus will be less noisy. This gives them trouble in standing out, and the resulting distribution underrepresents the peripheral candidates while giving equal probability to all the medial ones, as in the downward-pointing "ears" shown in (18).
  - In constraint-granularity NHG, all the candidates get same noise, N_1 and N_2, except the peripheral ones (maximally faithful, maximally unmarked). These are exempted from one noise factor and thus can stand out when this factor is high. The resulting distribution is rectilinear in the middle with upward-pointing "ears" at the periphery, as in (18).
  - Similar outcomes arise when we let violation-free cells have noise, but limit the noise to positive values.

3.5 *Summary of the modeling results*

I. To repeat the results of earlier researchers, Noisy Harmonic Grammar preserves the basic OT property of excluding harmonically bounded candidates, but only under certain conditions: noise must be constraint-specific, and perturbed weights must not allowed to go negative.

II. Various theories yield quite different behaviors for the local-optionality problem. Classical NHG respects harmonic bounding and awards probability only to the peripheral "lockstep" candidates. Maxent, and some versions of cell-specific NHG, yield rectilinear distributions; versions of NHG that multiply noise by violations are generally upsilonic; and versions that give special treatment to the peripheral candidates (either by assigning them no noise, or by forbidding negative weights) are rectilinear but with upward or downward tweaks at the periphery.

[10] www.linguistics.ucla.edu/people/hayes/VarietiesOfNHG. For convenience I have also included the predictions of the theories for Jesney's (2007) CVCC typology.

III. When a scalar constraint is pitted against a constant one, then pre-multiplicative noise (as in classical NHG), gives asymmetrical sigmoids with a nonzero asymptote. Post-multiplicative noise, and maxent, derive symmetrical sigmoids with asymptotes at zero and one.

4. The empirical side

The primary purpose of this paper has been to note differing predictions among stochastic constraint-based frameworks. But in the longer run these predictions should be put to the test, and I here summarize the few facts I am aware of that can bear on the differences pointed out here.

4.1 *The harmonic bounding controversy* The issue of whether harmonically bounded candidadtes can ever win seems to be a really fundamental difference between theories — and one that is hard to adjudicate empirically. It is often possible, as Kaplan (2016) showed, to remove instances of putative harmonic bounding by finding constraints violated by the bounding candidate. These constraints may well have their own empirical or theoretical motivation. However, it may be possible to argue for the existence of harmonically bounded candidates with nonzero probability on other grounds, as I will now suggest.

4.1.1 *Harmonically-bounded winners as the basis of gradient inventory theory* Theories that impose classical harmonic bounding seem a priori more restrictive, yet grammars that impose only the statistical version of harmonic bounding have arguable advantages. Specifically, there are quite a few analytic applications in linguistics where the goal is not to derive B from A, but to define a set of well-formed representations: grammatical sentences, morphologically-legal words, lines of verse, or words construed as phoneme strings. For all of these, membership in a putative "list of legal forms" is ill-defined, because native speakers' intuitions of well-formedness are gradient. A sensible strategy here is to create a grammar that assigns a probability value to every member of GEN (even harmonically bounded ones), with gradient well-formedness related to the probability of a candidate in some monotonic way. This has been done for phonotactics by Hayes and Wilson (2008) and for metrics by Hayes, Wilson, and Shisko (2012).

This procedure should be compared with the well-established analytic strategy for determining inventories in classical OT: the theory of the Rich Base, in which we carry out "pseudo-derivations" to underlie inventories — the legal forms are whatever can emerge from at least one input form (Prince and Smolensky 1993, Keer and Baković 1997, Hayes 2004, Prince and Tesar 2004, etc.). I think that the Rich Base theory is difficult to extend to gradience in well-formedness, because it predicts that semi-bad forms should often be repaired by speakers when they say them. What actually happens is that they are usually pronounced successfully, and their imperfection lies solely in the speaker's intuition, or the speaker's expectations in speech perception (Massaro and Cohen 1983, Moreton and Amano 1999).

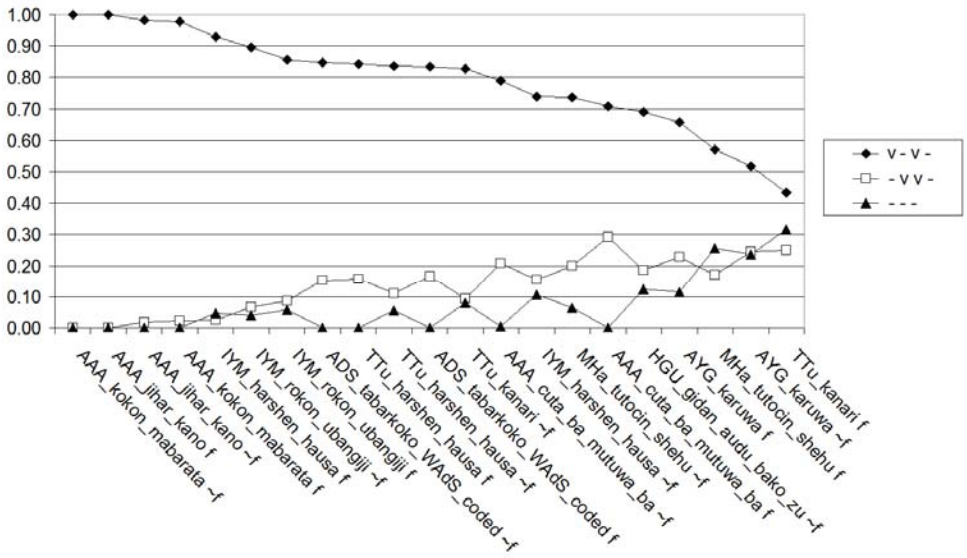
4.1.2 *An empirical example of gradient harmonic bounding* As noted above, various of theories described here impose a gradient (probability-difference) version of harmonic bounding. This prediction has not been checked out very widely, but I offer here an example from metrics, taken from Hayes and Schuh (in progress), who study the quantitative *rajaz* meter of Hausa. The meter is fundamentally iambic, based on repeated *light-heavy*; i.e. ⏑ – sequences, but there is a great deal of flexibility, and the "metron" (a sequence of two iambs) can be realized by various other hexamoraic patterns. In the Hayes/Schuh analysis, these other patterns violate prominence alignment constraints that affiliate syllable weight to metrical strength, and the violations form a pattern of harmonic bounding, as shown in (19).

(19) *A pattern of harmonic bounding in Hausa rajaz meter*

	STRONG POSITIONS MUST INITIATE HEAVY SYLLABLE	HEAVY SYLLABLE MUST INITIATE IN STRONG POSITION
x x x x x x x x ⏑ – ⏑ –		
x x x x x x x x – ⏑ ⏑ –	*	*
x x x x x x x x – – –	*	**

Hayes and Schuh checked the relative frequencies of these metra, for eleven different poems by seven different poets, and counting metra from stanza-final metra separately (since they generally behave differently). What emerged was a near-perfect pattern: no matter what the actual frequencies, the *relative* frequencies respected harmonic bounding relations for various poets/poems/stanza positions, as chart (20) shows.

(20) *Relative frequencies of three rajaz metron types across poets and stanza positions*



The few cases in which the inequalities [⏑ – ⏑ –] > [– ⏑ ⏑ –] and [– ⏑ ⏑ –] > [– – –] fail to hold involved relatively few data and could thus reflect accidents of sampling. The key idea of the analysis is that if all the poets are consistently using the same two constraints, and the harmonic bounding pattern of tableau (19) holds, then these relative frequency differences are inevitable. The example suggests a general type of case to look for; i.e. when the same constraints occurring in similar grammars repeatedly dictate a particular quantitative pattern among candidates related by harmonic bounding.

4.1.3 *The Goldrick/Daland theory of speech errors* This also depends on harmonically bounded winners; see discussion in §3.1.

4.1.4 *Multiple locus cases* These are, of course, a classic dilemma for the OT claim that harmonically bounded candidates always lose (Vaux 2003). To be sure, there are other solutions to the multiple-locus problem, cited above in §3.2. However, I think there is something appealing about getting multiple-locus cases — including, ideally, their fine-grained statistical detail — from the fundamental architecture of the theory, i.e. how probability distributions are computed.

4.2 *The characteristic shape of sigmoids in grammar* In principle it should be quite possible to check out the scenario of §3.3 empirically, examining the frequency curves generated by conflict between a scalar and an opposing constant constraint. McPherson and Hayes (2016:156) this did this for Tommo So vowel harmony, and found that using classical NHG (with asymmetrical sigmoids) obtained a somewhat inferior fit to the data; maxent and NHG with postmultiplicative noise worked best. The reason is what might be expected: the empirical data curves were symmetrical and (where the range of cases permitted this to be checked) asymptoted at zero and one.

The phonological world is, I believe, well endowed with sigmoids that are symmetrical and asymptote at zero and one; see in particular Zuraw and Hayes (in press). Indeed, the same is likely true of syntax (Kroch 1989, Bresnan et al. 2007), speech perception, and perhaps cognition in general. I think it would be of some interest if examples in any domain were located that displayed the asymmetrical, non-zero-asymptote pattern predicted by versions of NHG that multiply noise by constraint weights.

4.3 *Upsilonism vs. rectilinearism* I know of no data at all that bear on the question of upsilonism vs. rectilinearism. Perhaps it is experimentally testable. A difficulty is that free variation is responsive to style, and style is likely to create upsilonism independently of the constraint based frameworks: typically informal speaking style favors process application (at all loci) and formal style disfavors it. [11] To the extent that the data includes a mix of speaking styles, it will likely be upsilonic no matter what.

5. Conclusions

The empirical picture is patchy, but perhaps it will not remain so forever. In truth, the theoretical picture is likewise very incomplete: as far as locating differences of prediction among stochastic frameworks, the cases mentioned or cited here surely only scratch the surface.

References

Anttila, Arto (1997) Deriving variation from grammar. In Frans Hinskens, Roeland van Hout and W. Leo Wetzels (eds.), *Variation, change, and phonological theory*, 35–68. Amsterdam: John Benjamins.

Bod, Rens, Jennifer Hay and Stephanie Jannedy, eds. (2003) *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

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

Boersma, Paul and Joe Pater (2008/2016) Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John J. McCarthy and Joe Pater, eds., *Harmonic Grammar and Harmonic Serialism*. Sheffield: Equinox, pp. 389–434. [Original electronic distribution 2008.]

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen (2007) Predicting the dative alternation. In *Cognitive Foundations of Interpretation*, ed. by G. Boume, I. Kraemer, and J. Zwarts. Amsterdam: Royal Netherlands Academy of Science, pp. 69–94.

Coetzee, Andries (2004) What it means to be a loser: Non-optimal Candidates in Optimality Theory. Ph.D. dissertation, University of Massachusetts, Amherst.

Coetzee, Andries (2016) A comprehensive model of phonological variation: Grammatical and non-grammatical factors in variable nasal place assimilation. *Phonology* 33:211–246.

Coetzee, Andries and Shigeto Kawahara (2013) Frequency biases in phonological variation. *Natural Language and Linguisic Theory* 31:47–89.

[11] For recent treatment in Harmonic Grammar of factors that, like style, act as pan-phonology "knobs," see Coetzee and Kawahara (2013) and Coetzee (2016).

Fanselow, Gisbert, Caroline Féry, Ralf Vogel, and Matthias Schlesewsky, eds. (2006) *Gradience in grammar: Generative perspectives*. Oxford: Oxford University Press.

Goldrick, Matt and Robert Daland (2009) Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology* 26: 147–185.

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 Osten Dahl, 111–120.

Hayes, Bruce (2004). Phonological acquisition in Optimality Theory: the early stages. In René Kager, Joe Pater, and Wim Zonneveld (eds.), *Fixing priorities: constraints in phonological acquisition*. Cambridge University Press, 158–203.

Hayes, Bruce and Claire Moore-Cantwell (2011). Gerard Manley Hopkins's sprung rhythm: corpus study and stochastic grammar. *Phonology* 28:235–282

Hayes, Bruce and Russell Schuh (in progress) Metrical structure and sung rhythm of the Hausa rajaz. Ms., Department of Linguistics, UCLA.

Hayes, Bruce and Colin Wilson (2008) A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379–440.

Hayes, Bruce, Colin Wilson and Anne Shisko (2012) Maxent grammars for the metrics of Shakespeare and Milton, *Language* 88:691–731.

Hayes, Bruce, Kie Zuraw, Peter Siptár and Zsuzsa Londe (2009) Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85:822–863.

Hsu, Brian and Karen Jesney (2016) Scalar positional markedness and faithfulness in Harmonic Grammar. In *Proceedings of the Annual Meeting of the Chicago Linguistic Society* 51, 241–255. Chicago, IL: Chicago Linguistic Society.

Jäger, Gerhard (2007) Maximum entropy models and stochastic Optimality Theory. Architectures, rules, and preferences. *Variations on themes by Joan W. Bresnan*, ed. by Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling, and Chris Manning, 467–479. Stanford: CSLI Publications.

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

Kaplan, Aaron F. (2011) Variation through markedness suppression. *Phonology* 28:331–370.

Kaplan, Aaron F. (2016) Local optionality with partial orders. *Phonology* 33:285–324.

Keer, Edward and Eric Baković (1997) Have FAITH in syntax. In Emily Curtis, James Lyle, and Gabriel Webster, eds., *Proceedings of the Sixteenth West Coast Conference on Formal Linguistics*, pp. 255–269. Stanford, CA: Center for the Study of Language and Information.

Kimper, Wendell A. (2011) Locality and globality in phonological variation. *Natural Language and Linguistic Theory* 29:423–465.

Kiparsky, Paul (1982) Lexical phonology and morphology. In I. S. Yang (ed.), *Linguistics in the Morning Calm.* Seoul: Hanshin. 3–91.

Kroch, Anthony (1989) Reflexes of grammar in patterns of language change. *Language Variation and Change* 1:199–244.

Legendre, Geraldine, Yoshiro Miyata and Paul Smolensky (1990) Harmonic Grammar – A formal multi-level connectionist theory of linguistic well-formedness: An application. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 884–891. Mahwah, NJ: Lawrence Erlbaum Associates.

Legendre, Géraldine, Antonella Sorace and Paul Smolensky (2006) The Optimality Theory–Harmonic Grammar connection. In Paul Smolensky and Géraldine Legendre (eds.), *The Harmonic Mind*, 339–402. Cambridge, MA: MIT Press.

Massaro, Dominic and Michael M. Cohen (1983) Phonological context in speech perception. *Perception and Psychophysics* 34:338–348.

McPherson, Laura and Bruce Hayes (2016) Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology* 33:125–167.

Moreton, Elliott, and Shigeaki Amano (1999) Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies. *Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest*.

Nagy, Naomi and William Reynolds (1997) Optimality Theory and variable word-final deletion in Faetar. *Language variation and change* 9:37–55.

Osborn, Henry A. (1966) Warao I: Phonology and morphophonemics. *International Journal of American Linguistics* 32:108–123.

Pater, Joe (2008) Gradual learning and convergence. *Linguistic Inquiry* 39:334–345.

Pater, Joe (2008/2016) Universal grammar with weighted constraints. In John J. McCarthy and Joe Pater, eds., *Harmonic Grammar and Harmonic Serialism*. Sheffield: Equinox, pp. 1–46. [Original electronic distribution 2008.]

Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt and Michael Becker (2010) Harmonic Grammar with linear programming: From linear systems to linguistic typology. *Phonology* 27:77–117.

Prince, Alan (2002) Arguing optimality. Rutgers Optimality Archive 562.

Prince, Alan and Paul Smolensky (1993/2004) *Optimality Theory: Constraint interaction in generative grammar*. Technical report, Rutgers University Center for Cognitive Science. [Published 2004; Oxford: Blackwell.]

Prince, Alan, and Bruce Tesar (2004) Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 245–291. Cambridge: Cambridge University Press.

Riggle, Jason and Colin Wilson (2005) Local optionality. In Leah Bateman and Cherlon Ussery (eds.), *Proceedings of the North Eastern Linguistic Society* 35.

Samek-Lodovici, Vieri and Alan Prince (2002) Fundamental properties of harmonic bounding. RuCCS-TR-71.

Smolensky, Paul (1986) Information processing in dynamical systems: foundations of Harmony Theory. In *Parallel distributed processing: explorations* in *the microstructure of cognition*, ed. David E. Rumelhart, James L. McClelland, and the PDP Research Group, volume 1, 194–281. Cambridge, Mass.: MIT Press/Bradford Books.

Vaux, Bert (2003) Why the phonological component must be serial and rule-based. Paper presented at the Linguistic Society of America.

Vaux, Bert (2008) Why the phonological component must be serial and rule-based. In Bert Vaux and Andrew Nevins, eds., *Rules, constraints, and phonological phenomena*. Oxford: Oxford University Press.

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

Wilson, Colin, and Benjamin George (2009) Maxent grammar tool. Software. http://www.linguistics.ucla.edu/people/hayes/MaxentGrammarTool/

Zuraw, Kie (2000) Patterned exceptions in phonology. Ph.D. dissertation, UCLA, Los Angeles, CA.

Zuraw, Kie (2010) A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28:417–472.

Zuraw, Kie, and Bruce Hayes (in press) Intersecting constraint families: an argument for Harmonic Grammar. To appear in *Language*.