Stochastic harmonic grammars do not peak on the mappings singled out by categorical harmonic grammars

Giorgio Magri
CNRS, MIT
magrigrg@gmail.com

Abstract

A candidate surface phonological realization is called a peak of a probabilistic constraint-based phonological grammar provided it achieves the largest probability mass over its candidate set. Obviously, the set of peaks of a maximum entropy grammar is the categorical harmonic grammar corresponding to the same weights. This paper shows that the set of peaks of a stochastic harmonic grammar instead can be different from the categorical harmonic grammar corresponding to any weights. Thus in particular, maximum entropy and stochastic harmonic grammars can peak on different candidates.

Maximum Entropy grammars (ME; Goldwater and Johnson, 2003; Hayes and Wilson, 2008) and Noisy or Stochastic Harmonic Grammars (SHG; Boersma and Pater, 2016) are both probabilistic extensions of categorical Harmonic Grammars (HG; Legendre et al., 1990b,a; Pater, 2009). A growing body of literature tries to pull apart these two probabilistic frameworks. One line of research compares ME and SHG in terms of their ability to fit specific patterns of data given specific choices of candidates and constraints (Zuraw and Hayes, 2017; Smith and Pater, 2020; Breiss and Albright, 2022). Another line of research compares their typological predictions independently of the choice of the constraints, by characterizing the uniform probability inequalities they predict (Anttila and Magri, 2018; Anttila et al., 2019; Magri and Anttila, 2023).

This paper compares ME and SHG in terms of their probability peaks, namely the candidates to which they assign largest probability masses, as formalized in section 1. Obviously, the peaks of the ME grammar corresponding to some non-negative weights are the winners singled out by the categorical HG grammar corresponding to the same weights, no matter what the constraint set looks like, as illustrated in section 2. In other words, ME grammars peak on HG winners. Crucially, this property does not extend from ME to SHG, as discussed in section 3. Indeed, section 4 constructs an example of SHG grammar whose peaks cannot be described as the HG winners corresponding to any non-negative weights. In other words, SHG grammars do not necessarily peak on HG winners. It follows in particular that ME and SHG grammars can peak on different candidates.

The proposed counterexample is somewhat contrived and no simpler counterexamples seem readily available. It is therefore improbable that we would ever "stumble" into one such counterexample by simply "playing" with SHG phonology. This result about SHG peaks thus shows that only mathematical analysis can reveal subtle properties of probabilistic phonological models—which is one of the main goals of linguistic theory.

1 Peaks of probabilistic grammars

A phonological mapping is a pair (x, y) consisting of an underlying form x and a corresponding surface realization y. Gen denotes the set of mappings relevant for the description of the phonological system of interest (Prince and Smolensky, 1993/2004). $Gen(\mathsf{x})$ denotes the set of candidate surface realizations y such that the mapping (x, y) belongs to Gen. We allow Gen to list countably infinitely many underlying forms. But we require a candidate set $Gen(\mathsf{x})$ to be finite to avoid the technicalities needed to define probability mass functions on infinite sets.

A **categorical grammar** G assigns to an underlying form x a unique "winner" surface realization y from the candidate set $Gen(\mathsf{x})$. Thus, we require categorical grammars to be **strict**: they specify a unique surface realization per underlying form. On the other hand, we allow categorical grammars to be **partial**: they might fail to specify any surface realization for a given underlying form. HG grammars recalled below are indeed usually defined as strict and partial.

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 269-277.
Amherst, Massachusetts, June 15-17, 2023

A **probabilistic grammar** G assigns to each mapping (x, y) listed by Gen a number $G(\mathsf{y} \,|\, \mathsf{x})$ that is interpreted as the probability that the underlying form x is realized as the surface candidate y. This probabilistic interpretation requires these numbers $G(\mathsf{y} \,|\, \mathsf{x})$ to be non-negative and normalized across the candidate set $Gen(\mathsf{x})$ of each underlying form x, namely $\sum_{\mathsf{y} \in Gen(\mathsf{x})} G(\mathsf{y} \,|\, \mathsf{x}) = 1$.

We say that a mapping (x, y) is a **peak** of a probabilistic grammar G provided y is assigned a larger probability mass than any other candidate z of the underlying form x, as stated in (1).

$$G(\mathsf{y} \,|\, \mathsf{x}) > \max_{\substack{\mathsf{z} \in Gen(\mathsf{x}) \\ \mathsf{z} \neq \mathsf{y}}} G(\mathsf{z} \,|\, \mathsf{x}) \tag{1}$$

The set of candidates with peak probabilities can be interpreted as a categorical grammar. This categorical grammar is strict, because condition (1) features a strict inequality, whereby at most one candidate per underlying form qualifies as a peak. Furthermore, this categorical grammar is partial, because multiple candidates can tie for the largest probability, whereby none qualifies as a peak.

Intuitively, these candidates that are assigned the largest probability masses are those that are deemed most important by a probabilistic grammar. The set of these most important candidates with peak probabilities thus ought to capture some important information about the probabilistic grammar. As a first stub at analyzing a complex probabilistic grammar, it thus makes sense to analyze the corresponding categorical grammar of peaks.

2 ME peaks are HG winners

To illustrate the definitions in the preceding section, we consider a set $\boldsymbol{C}$ consisting of a finite number n of constraints C_k. We denote by $C_k(\mathsf{x}, \mathsf{y})$ the number of violations assigned by constraint C_k to a mapping (x, y) from Gen. We assign to each constraint C_k a non-negative weight w_k. A candidate y is the winner surface realization of an underlying form x provided it satisfies condition (2). It says that the candidate y violates the constraints less than any other candidate z because the weighted sum of the constraint violations of y is strictly smaller. The categorical grammar G that singles out such winner candidates is the **HG** grammar corresponding to the weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)$. It is strict, because (2) features a strict inequality. It can be partial, in case two or more candidates tie for the smallest weighted sum of constraint violations.

$$\sum_{k=1}^{n} w_k C_k(\mathsf{x}, \mathsf{y}) < \min_{\substack{\mathsf{z} \in Gen(\mathsf{x}) \\ \mathsf{z} \neq \mathsf{y}}} \sum_{k=1}^{n} w_k C_k(\mathsf{x}, \mathsf{z}) \tag{2}$$

We can also use the constraint set $\boldsymbol{C}$ and the weight vector $\boldsymbol{w}$ to define a probabilistic grammar though condition (3). It says that the probability $G(\mathsf{y} \,|\, \mathsf{x})$ that an underlying form x is realized as a candidate y is the exponential of the opposite of the weighted sum of constraint violations of the mapping (x, y), divided by a quantity $Z(\mathsf{x})$ that ensures normalization over the candidate set $Gen(\mathsf{x})$. The resulting probabilistic grammar G is the **ME** grammar corresponding to the weight vector $\boldsymbol{w}$.

$$G(\mathsf{y} \,|\, \mathsf{x}) = \frac{1}{Z(\mathsf{x})} \exp\left\{ -\sum_{k=1}^{n} w_k C_k(\mathsf{x}, \mathsf{y}) \right\} \tag{3}$$

The normalization constant $Z(\mathsf{x})$ depends on the underlying form x but not on the candidate y. Furthermore, the definition (1) of probability peaks only compares probabilities within the same candidate set. It follows that a mapping (x, y) qualifies as a peak of the ME grammar (3) corresponding to the weight vector $\boldsymbol{w}$ if and only if (x, y) belongs to the HG grammar (2) corresponding to the same weight vector $\boldsymbol{w}$. In other words, HG grammars single out the peaks of ME grammars.

3 SHG peaks are not HG winners

Let p_w be a uni-dimensional probability density function that starts at a point w, in the sense that it is equal to zero at the left of w. Here are some natural examples of such a density ($\mathbb{I}_S$ is the indicator function of the set S):

- the uniform density $p_w^{\text{unif}}(v) = \mathbb{I}_{[w,w+1]}(v)$;
- the exponential density $p_w^{\text{exp}}(v) = \exp(w - v)\mathbb{I}_{[w,+\infty]}(v)$;
- the half-gaussian density $p_w^{\text{gauss}}(v) = \frac{2\exp\left[-(v-w)^2/2\right]}{\sqrt{2\pi}}\mathbb{I}_{[w,+\infty]}(v)$

Given a constraint set $\boldsymbol{C}$ and a non-negative weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)$, the corresponding **SHG** grammar assigns to a mapping (x, y) the probability of sampling a weight vector $\boldsymbol{v}$ according to $\mathbf{p}_{\boldsymbol{w}} = p_{w_1} \cdot \ldots \cdot p_{w_n}$ such that y qualifies as an HG winner

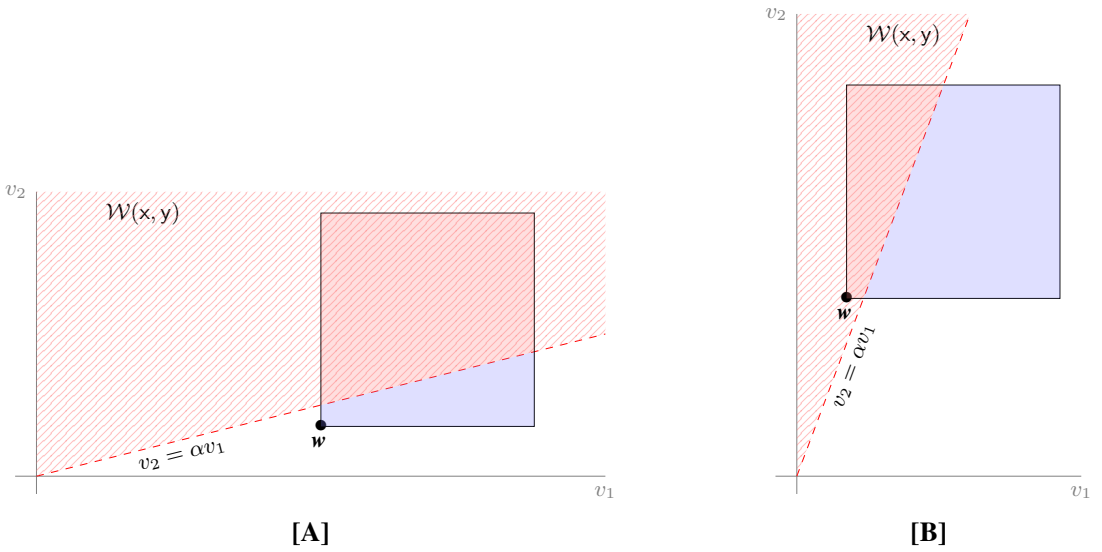


Figure 1

corresponding to this weight vector $\boldsymbol{v}$ in the sense of condition (2) above. The assumption that the weights w_k are non-negative and that p_{w_k} starts at w_k ensures that the probability of sampling vectors $\boldsymbol{v}$ with negative components is zero.[1]

To understand intuitively why SHG peaks are not necessarily HG winners, let us consider the following simplest case. *Gen* lists only two candidate surface realizations y and z for an underlying form x. The constraint set consists of only $n = 2$ constraints. The set $\mathcal{W}(\mathsf{x},\mathsf{y})$ of weight vectors $\boldsymbol{v} = (v_1, v_2)$ such that the weighted sum of constraint violations of y is smaller than that of z is the dashed red cone in figure 1 described by the inequality $v_2 > \alpha v_1$, for some $\alpha > 0$. The SHG grammar corresponding to a weight vector $\boldsymbol{w}$ is implemented with the uniform density that concentrates the probability mass on the square that starts at the weight vector $\boldsymbol{w}$. This square is split into two halves by the red dashed cone $\mathcal{W}(\mathsf{x},\mathsf{y})$. The area of the solid red half of the square that sits within the cone $\mathcal{W}(\mathsf{x},\mathsf{y})$ is the probability mass assigned by our SHG grammar to the mapping (x,y). The area of the remaining solid blue half that sits outside of $\mathcal{W}(\mathsf{x},\mathsf{y})$ is the probability mass assigned to (x,z).

The weight vector $\boldsymbol{w}$ in figure 1A sits outside of the cone $\mathcal{W}(\mathsf{x},\mathsf{y})$. Thus, the HG grammar corresponding to $\boldsymbol{w}$ does not contain the mapping (x,y). Yet, $\boldsymbol{w}$ sits so close to the border of the cone $\mathcal{W}(\mathsf{x},\mathsf{y})$ that the red solid area is larger than the blue solid area. Thus, our mapping (x,y) is a peak of the SHG grammar corresponding to the weight vector $\boldsymbol{w}$, because (x,y) receives a larger probability mass than (x,z). In conclusion, a peak of an SHG grammar might not belong to the HG grammar corresponding to the same weight vector.

Figure 1B illustrates the reverse scenario. The HG grammar contains the mapping (x,y) because the weight vector $\boldsymbol{w}$ sits inside the cone $\mathcal{W}(\mathsf{x},\mathsf{y})$. Yet, $\boldsymbol{w}$ sits so close to the border of the cone that the mapping (x,y) does not count as a peak of the SHG grammar because the red solid area is smaller than the blue solid area. In conclusion, a mapping of an HG grammar might not be a peak of the SHG grammar corresponding to the same weight vector.

These mismatches between SHG peaks and HG mappings are only possible when the border of the cone $\mathcal{W}(\mathsf{x},\mathsf{y})$ is **less** tilted than the diagonal because $\alpha < 1$ as in figure 1A; or it is **more** titled than the diagonal because $\alpha > 1$ as in figure 1B. These mismatches are not possible when instead the border of the cone $\mathcal{W}(\mathsf{x},\mathsf{y})$ coincides with the

[1] The implementation of probabilistic constraint-based phonology that I call here "stochastic" HG is slightly different from what Boersma and Pater (2016) call "noisy" HG, because the two implementations differ for the strategy they adopt to avoid sampling zero weights. In SHG, zero weights are avoided by sampling according to a density p_w that starts at a positive value w. In NHG, zero weights are avoided by clipping at zero or by re-sampling (Hayes and Kaplan 2023). Furthermore, the term "stochastic" HG makes it explicit that the resulting framework is a probabilistic extension of categorical HG in exactly the same way that Stochastic OT is a probabilistic extension of categorical OT (Boersma 1997, 1998). Finally, the term "noisy" is traditionally used to qualify the training data, while "stochastic" is used to single out algorithms (and thus grammars) that are non-deterministic.

diagonal because $\alpha = 1$. In this case, the red solid area is larger (smaller) than the blue solid area if and only if the weight vector $\boldsymbol{w}$ sits inside (outside) of the cone $\mathcal{W}(\mathsf{x}, \mathsf{y})$, no matter how close $\boldsymbol{w}$ is to the diagonal border of the cone. As a result, (x, y) is an SHG peak if and only if it belongs to the HG grammar corresponding to the same weights. We will use this observation in subsection 4.1 below.

The considerations developed so far for the uniform density based on elementary geometric considerations extend to other densities. To illustrate, let us consider the exponential density. We start with the case where the border of the cone $\mathcal{W}(\mathsf{x}, \mathsf{y})$ is **less tilted** than the diagonal as in figure 1A, say because $\alpha = 1/2$. We focus on weight vectors $\boldsymbol{w} = (w_1, w_2)$ that sit outside of this cone because they have a negative "distance" $\xi = w_2 - \alpha w_1 < 0$ from the border of the cone. The SHG probability mass of our mapping (x, y) is easily computed in closed form by integrating the exponential function. Figure 2A plots this SHG probability mass (on the vertical axis) as a function of the "distance" $\xi < 0$ (on the horizontal axis). When ξ is between $\log(3/4) \simeq -0.2877$ and zero, the weight vector $\boldsymbol{w}$ sits outside of the cone $\mathcal{W}(\mathsf{x}, \mathsf{y})$, whereby the mapping (x, y) does not belong to the corresponding HG grammar. Yet (x, y) is a peak of the corresponding SHG grammar, because the SHG probability mass of y is larger than 0.5, and therefore larger than the SHG probability mass of z.

Analogously, let us consider the case where the border of the cone $\mathcal{W}(\mathsf{x}, \mathsf{y})$ is **more tilted** than the diagonal as in figure 1B, say because $\alpha = 2$. We focus on weight vectors $\boldsymbol{w} = (w_1, w_2)$ that sit inside this cone because they have a positive "distance" $\xi = w_2 - \alpha w_1 > 0$ from the border of the cone. Figure 2B plots the SHG probability mass of our mapping (x, y) as a function of the "distance" $\xi > 0$. When ξ is between zero and $\log(16/9) \simeq 0.5754$, the weight vector $\boldsymbol{w}$ sits inside the cone $\mathcal{W}(\mathsf{x}, \mathsf{y})$, whereby the mapping (x, y) does belong to the corresponding HG grammar. Yet (x, y) is not a peak of the corresponding SHG grammar, because the SHG probability mass of y is smaller than 0.5, and therefore smaller than the SHG probability mass of z.

These considerations show that the mappings singled out by the HG grammar corresponding to some weight vector $\boldsymbol{w}$ are not necessarily the peaks of the SHG grammar corresponding to the same weight vector $\boldsymbol{w}$. Yet, it can be shown (through a different line of analysis that falls outside of the scope of this paper), that the mappings singled out by the HG grammar corresponding to some weight vector $\boldsymbol{w}$ are always the peaks of the SHG grammar corresponding to a possibly different weight vector $\boldsymbol{w}'$. What about the reverse? Despite the mismatches between SHG peaks and HG mappings documented above, is it the case that the peaks of the SHG grammar corresponding to some weight vector $\boldsymbol{w}$ are always the mappings singled out by the HG grammar corresponding to a possibly different weight vector $\boldsymbol{w}'$? The next section provides a negative answer to this question by constructing an SHG grammar whose set of peaks is not an HG grammar, no matter the choice of the weights.

4 Counterexample

To construct the simplest possible counterexample, we assume that Gen lists only three underlying forms x_1, x_2, and x_3 and endows each of them with only two candidates y_i and z_i, as in (4)

$$Gen = \left\{ \begin{array}{ccc} (\mathsf{x}_1, \mathsf{y}_1) & (\mathsf{x}_2, \mathsf{y}_2) & (\mathsf{x}_3, \mathsf{y}_3) \\ (\mathsf{x}_1, \mathsf{z}_1) & (\mathsf{x}_2, \mathsf{z}_2) & (\mathsf{x}_3, \mathsf{z}_3) \end{array} \right\} \quad (4)$$

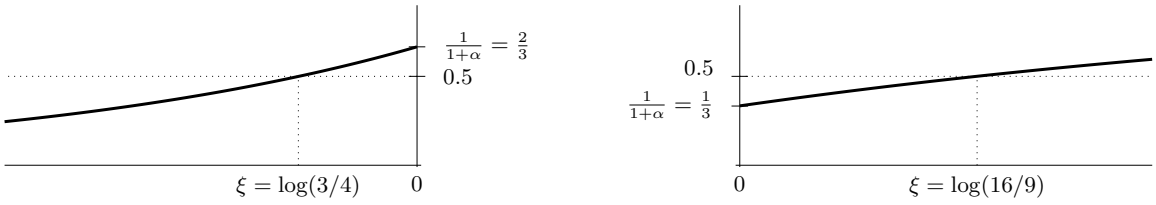
The constraint set $\boldsymbol{C}$ consists of $n = 3$ constraints C_1, C_2, and C_3 that yield the violation profiles in (5). Actual numbers of constraint violations do not matter. What does matter for the counterexample are the ratios of the differences between the numbers of violations of two candidates, as shown in appendix E. To illustrate, it does not matter that C_1 and C_3 assign 33 and 0 violations to y_3 and 0 and 200 violations to z_3. What does matter is that the ratio between the differences $C_1(\mathsf{x}_3, \mathsf{y}_3) - C_1(\mathsf{x}_3, \mathsf{z}_3)$ and $C_3(\mathsf{x}_3, \mathsf{z}_3) - C_3(\mathsf{x}_3, \mathsf{y}_3)$ is equal to $33/200 = 0.165$. These large numbers 33 and 200 are needed to express a small value 0.165 as the ratio 33/200 between two integers.

	C_1	C_2	C_3
$(\mathsf{x}_1, \mathsf{y}_1)$	0	5	0
$(\mathsf{x}_1, \mathsf{z}_1)$	2	0	0
$(\mathsf{x}_2, \mathsf{y}_2)$	0	0	5
$(\mathsf{x}_2, \mathsf{z}_2)$	0	2	0
$(\mathsf{x}_3, \mathsf{y}_3)$	33	0	0
$(\mathsf{x}_3, \mathsf{z}_3)$	0	0	200

(5)

Finally, the vector $\boldsymbol{w} = (w_1, w_2, w_3)$ of non-negative weights is chosen carefully as in (6).

$$\begin{array}{rcl} w_1 & = & 4.21734890439 \\ w_2 & = & 1.3195643695 \\ w_3 & = & 0.160450555542 \end{array} \quad (6)$$



[A] Case $\alpha = 1/2$ **[B]** Case $\alpha = 2$

Figure 2

We implement the SHG grammar with the exponential density. When this SHG grammar is called for 100,000 times on each of the three underlying forms x_1, x_2, and x_3, it returns the surface forms y_1, y_2, and y_3 with the frequencies in (7).

$$\begin{aligned} G_{\boldsymbol{w}}^{\text{SHG}}(\mathsf{y}_1 \,|\, \mathsf{x}_1) &= 0.50566 \\ G_{\boldsymbol{w}}^{\text{SHG}}(\mathsf{y}_2 \,|\, \mathsf{x}_2) &= 0.50637 \\ G_{\boldsymbol{w}}^{\text{SHG}}(\mathsf{y}_3 \,|\, \mathsf{x}_3) &= 0.50126 \end{aligned} \tag{7}$$

Since these frequencies are (close to but strictly) larger than 0.5, the categorical grammar of peaks of the SHG grammar considered is the grammar G in (8). Crucially, we will see that this grammar G is not an HG grammar corresponding to any choice of non-negative constraint weights.

$$G = \big\{(\mathsf{x}_1, \mathsf{y}_1), (\mathsf{x}_2, \mathsf{y}_2), (\mathsf{x}_3, \mathsf{y}_3)\big\} \tag{8}$$

In conclusion, we have constructed an SHG grammars whose set of peaks cannot be construed as any HG grammar. The rest of this section explains in detail how the counterexample has been constructed, by motivating the choice of the violation profiles in (5) and of the weights in (6).

4.1 First step

We need to define the $n = 3$ constraints in such a way that the grammar G in (8) is not an HG grammar. As above, let us denote by $\mathcal{W}(\mathsf{x}_i, \mathsf{y}_i)$ the cone of those non-negative weight vectors $\boldsymbol{v} = (v_1, v_2, v_3)$ that declare y_i the winner surface realization of the underlying form x_i. A simple strategy to achieve our goal is to define the constraints so that these three cones are as in (9). In fact, the grammar G in (8) qualifies as an HG grammar only if some non-negative weight vector $\boldsymbol{v} = (v_1, v_2, v_3)$ belongs simultaneously to all three cones. And that is impossible. Because a weight vector that belongs to both cones $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ and $\mathcal{W}(\mathsf{x}_2, \mathsf{y}_2)$ must satisfy both inequalities $v_1 > v_2$ and $v_2 > v_3$. By transitivity, it must also satisfy the inequality $v_1 > v_3$. Hence, our weight vector cannot belong to the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$.

$$\begin{aligned} \mathcal{W}(\mathsf{x}_1, \mathsf{y}_1) &= \big\{\boldsymbol{v} \,\big|\, v_1 > v_2\big\} \\ \mathcal{W}(\mathsf{x}_2, \mathsf{y}_2) &= \big\{\boldsymbol{v} \,\big|\, v_2 > v_3\big\} \\ \mathcal{W}(\mathsf{x}_3, \mathsf{y}_3) &= \big\{\boldsymbol{v} \,\big|\, v_1 < v_3\big\} \end{aligned} \tag{9}$$

Unfortunately, the borders of the cones in (9) are all diagonal. As discussed in section 3 above, we get no mismatches between SHG peaks and HG mappings in this case. Thus, I make the borders non-diagonal by replacing the cones in (9) with those in (10), where the steepness of the borders is controlled by the positive coefficients a and α. I use the same coefficient a for both cones $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ and $\mathcal{W}(\mathsf{x}_2, \mathsf{y}_2)$, as this choice simplifies the analysis without compromising the counterexample.

$$\begin{aligned} \mathcal{W}(\mathsf{x}_1, \mathsf{y}_1) &= \big\{\boldsymbol{v} \,\big|\, v_1 > a\, v_2\big\} \\ \mathcal{W}(\mathsf{x}_2, \mathsf{y}_2) &= \big\{\boldsymbol{v} \,\big|\, v_2 > a\, v_3\big\} \\ \mathcal{W}(\mathsf{x}_3, \mathsf{y}_3) &= \big\{\boldsymbol{v} \,\big|\, v_3 > \alpha\, v_1\big\} \end{aligned} \tag{10}$$

As for the HG-hood of the grammar G in (8), the replacement of our initial guess (9) with the refined guess (10) changes nothing because of the following lemma, verified in appendix A.

Lemma 1 *Suppose that the positive coefficients $a, \alpha > 0$ satisfy condition (11).*

$$a^2 \alpha \geq 1 \tag{11}$$

No weight vector $\boldsymbol{v} = (v_1, v_2, v_3)$ belongs simultaneously to the three cones in (10), whereby the grammar G in (8) is not an HG grammar.

4.2 Second step

We now want to construct a non-negative weight vector $\boldsymbol{w} = (w_1, w_2, w_3)$ such that the peaks of the corresponding SHG grammar are indeed the three mappings singled out by the grammar G in (8). As discussed in the preceding subsection, this weight vector $\boldsymbol{w}$ cannot belong simultaneously to all three cones in (10). For concreteness, we assume that the

weight vector $\boldsymbol{w} = (w_1, w_2, w_3)$ does not belong to the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ while it does belong to the other two cones $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ and $\mathcal{W}(\mathsf{x}_2, \mathsf{y}_2)$.

The assumption that the weight vector $\boldsymbol{w}$ sits outside of the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ despite the mapping $(\mathsf{x}_3, \mathsf{y}_3)$ being a peak of the corresponding SHG grammar has two consequences. The first consequence is that the border of the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ must be less tilted than the diagonal, as in figure 1A. In other words, the coefficient α that controls its tiltedness must be small in the sense that $\alpha < 1$. The second consequence is that, although the weight vector $\boldsymbol{w}$ sits outside of the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$, it cannot be too far away from it. Equivalently, although w_3 is smaller than αw_1 (so that $\boldsymbol{w}$ sits outside of $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$), it cannot be too much smaller (so that $\boldsymbol{w}$ sits close to $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$). Not much smaller in the sense that the weights w_1 and w_3 satisfy the inequality $w_3 > \alpha w_1 - A$ for some carefully chosen positive constant $A > 0$. The following lemma says that we need to choose this constant A equal to $\log \frac{2}{1+\alpha}$, as verified in appendix B. Since $\alpha < 1$, this position $A = \log \frac{2}{1+\alpha}$ is positive as desired.

Lemma 2 *Consider a weight vector* $\boldsymbol{w} = (w_1, w_2, w_3)$ *that does not belong to the cone* $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ *because* $w_3 < \alpha w_1$. *The mapping* $(\mathsf{x}_3, \mathsf{y}_3)$ *is a peak of the SHG grammar corresponding to this weight vector* $\boldsymbol{w}$ *provided* $\boldsymbol{w}$ *satisfies (12).*

$$w_3 > \alpha w_1 - \underbrace{\log \frac{2}{1+\alpha}}_{A} \tag{12}$$

Condition (11) together with the assumption $\alpha < 1$ made above entails that the coefficient a that controls the tiltedness of the border of the cone $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ is large in the sense that $a > 1$. Equivalently, the border of the cone $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ is steeper than the diagonal. As a result, the assumption that the weight vector $\boldsymbol{w}$ sits inside the cone $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ by itself does not suffice to ensure that $(\mathsf{x}_1, \mathsf{y}_1)$ is a peak, as shown in figure 1B. We need to make sure that the weight vector $\boldsymbol{w}$ sits well inside this cone $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$, far away from the border. Equivalently, w_1 is not just larger than aw_2 (so that $\boldsymbol{w}$ sits inside $\mathcal{W}(\mathsf{x}_1\mathsf{y}_1)$) but actually quite larger (so that $\boldsymbol{w}$ sits well inside $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$). Quite larger in the sense that the weights w_1 and w_2 satisfy the inequality $w_1 > aw_2 + B$ for some carefully chosen positive constant $B > 0$. The following lemma says that we need to choose this constant B equal to $a \log \frac{2a}{1+a}$, as verified in appendix C. Since $a > 1$, this position $B = a \log \frac{2a}{1+a}$ is positive as desired.

Lemma 3 *Consider a weight vector* $\boldsymbol{w} = (w_1, w_2, w_3)$ *that does belong to the cone* $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ *because* $w_1 > aw_2$. *The mapping* $(\mathsf{x}_1, \mathsf{y}_1)$ *is a peak of the SHG grammar corresponding to this weight vector* $\boldsymbol{w}$ *provided* $\boldsymbol{w}$ *satisfies (13).*

$$w_1 > aw_2 + \underbrace{a \log \frac{2a}{1+a}}_{B} \tag{13}$$

A completely analogous reasoning shows that condition (14) ensures that the mapping $(\mathsf{x}_2, \mathsf{y}_2)$ is a peak of the SHG grammar corresponding to the weight vector $\boldsymbol{w} = (w_1, w_2, w_3)$.

$$w_2 > aw_3 + a \log \frac{2a}{1+a} \tag{14}$$

4.3 Third step

Do the three inequalities (12), (13), and (14) just obtained admit non-negative solutions $w_1, w_2, w_3 \geq 0$? To answer this question, we use of the following straightforward fact, verified in appendix D.

Lemma 4 *Suppose that* $a^2\alpha > 1$, *as in (11). The following three strict inequalities*

$$\begin{array}{rcl} w_3 & > & \alpha w_1 - A \\ w_1 & > & aw_2 + B \\ w_2 & > & aw_3 + B \end{array} \tag{15}$$

admit non-negative solutions $w_1, w_2, w_3 \geq 0$ *when their coefficients* $a, \alpha > 0$ *and* $A, B \geq 0$ *satisfy the following condition (16).*

$$1 < \frac{A}{\alpha(1+a)B} \tag{16}$$

Indeed, the inequalities (12), (13), and (14) have the shape in (15) with the positions (17).

$$A = \log \frac{2}{1+\alpha}, \qquad B = a \log \frac{2a}{1+a} \tag{17}$$

Condition (16) that ensures that the three inequalities (15) admit non-negative solutions boils down to condition (18) with these positions (17). We conclude that the inequalities (12), (13), (14) admit non-negative solutions $w_1, w_2, w_3 \geq 0$ when the coefficients a, α satisfy condition (18).

$$\frac{1}{\alpha} \log \frac{2}{1+\alpha} - a(1+a) \log \frac{2a}{1+a} > 0 \tag{18}$$

4.4 Fourth step

In conclusion, in order for our counterexample to work, we need to find coefficients $a > 1$ and $0 < \alpha < 1$ that satisfy both conditions (11) and (18). To this end, figure 3 plots in red (blue) the pairs of values (a, α) that satisfy (do not satisfy) condition (18). Furthermore, the black line in figure 3 describes the equation $\alpha = 1/a^2$. The pairs of values (a, α) that satisfy condition (11) thus sit above and at the right of this black line. This figure thus says that a pair of values (a, α) satisfies both conditions (11) and (18) as desired provided it belongs to the narrow band between the black line and the boundary between the red and blue regions. The pair (a, α) in (19) belongs indeed to this narrow band and thus satisfies both conditions (11) and (18).

$$a = 2.5, \qquad \alpha = 0.165 \tag{19}$$

When the constraint violation vectors are defined as in (5), the cones $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$, $\mathcal{W}(\mathsf{x}_2, \mathsf{y}_2)$, and $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ are precisely the cones described by the inequalities in (10) with the coefficients a and α as in (19), because $5/2 = 2.5 = a$ and $33/200 = 0.165 = \alpha$, as verified in appendix E. Finally, the three inequalities (12), (13), and (14) corresponding to the coefficients a and α in (19) admit non-negative solutions w_1, w_2, w_3 such as those in (6), as shown in appendix F, completing the explanation of the counterexample.

5 Conclusions

Categorical grammars can usually be analyzed by exhaustive enumeration and direct inspection of the mappings they contain. Probabilistic grammars instead require more sophisticated analytical tools. A natural idea is to analyze some of the linguistic information captured by a complex probabilistic grammars by analyzing it peaks, namely the candidates that are deemed most important by that probabilistic grammar because assigned the largest probability mass. For a ME grammar, this is easily done: its peaks are the HG winners corresponding to the same weight vector. This paper has shown that the situation is different in SHG: although any HG grammar can be construed as the set of peaks of some SHG grammar, the set of peaks of some SHG grammars cannot be construed as an HG grammar, no matter the choice of the weights. It follows that ME and SHG grammars corresponding to the same weights can peak on different candidates.

Appendix

A Proof of lemma 1

A weight vector $\boldsymbol{v} = (v_1, v_2, v_3)$ that belongs to both cones $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$ and $\mathcal{W}(\mathsf{x}_2, \mathsf{y}_2)$ satisfies both inequalities $v_1 > av_2$ and $v_2 > av_3$. Thus in particular, $\boldsymbol{v}$ satisfies the inequality $v_1 > a^2 v_3$. On the other hand, a weight vector $\boldsymbol{v}$ that belongs to the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ satisfies the inequality $v_1 < v_3/\alpha$. These two inequalities yield $a^2 v_3 < v_3/\alpha$. Since this inequality is strict, v_3 must be strictly positive and can therefore be simplified, yielding $a^2 < \frac{1}{\alpha}$. This conclusion contradicts the assumption (11).

B Proof of lemma 2

We start by establishing the chain of identities in (20). Step (20a) below holds because of the definition of the exponential density. Step (20b) holds because $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ is the cone consisting of the non-negative vectors $\boldsymbol{v} = (v_1, v_2, v_3)$ such that $v_3 \geq \alpha v_1$. Step (20c) holds because of the hypothesis $w_3 \leq \alpha w_1$ that $\boldsymbol{w} = (w_1, w_2, w_3)$ sits outside of the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$. Thus, $v_1 \geq w_1$ entails $\alpha v_1 \geq \alpha w_1 \geq w_3$, whereby $\max\{w_3, \alpha v_1\} = \alpha v_1$. The remaining steps only use the identity $\int e^{-\lambda x} \mathrm{d}x = -\frac{1}{a} e^{-\lambda x}$.

$$\begin{aligned}
&\int_{\mathcal{W}(\mathsf{x}_3,\mathsf{y}_3)} p^{\exp}_{w_1}(v_1)\, p^{\exp}_{w_3}(v_3)\, \mathrm{d}v_1\, \mathrm{d}v_3 = \\
&\overset{(a)}{=} e^{w_1+w_3} \int_{v_1 \geq w_1, v_3 \geq w_3} e^{-v_1-v_3} \mathbb{I}_{\mathcal{W}(\mathsf{x},\mathsf{y})}(v_1, v_3) \mathrm{d}v_1 \mathrm{d}v_3 \\
&\overset{(b)}{=} e^{w_1} e^{w_3} \int_{v_1 \geq w_1} e^{-v_1} \int_{v_3 \geq \max\{w_3, \alpha v_1\}} e^{-v_3} \mathrm{d}v_1 \mathrm{d}v_3 \\
&\overset{(c)}{=} e^{w_1+w_3} \int_{v_1 \geq w_1} e^{-v_1} \int_{v_3 \geq \alpha v_1} e^{-v_3} \mathrm{d}v_3 \mathrm{d}v_1 \\
&= e^{w_1+w_3} \int_{v_1 \geq w_1} e^{-v_1} \left| -e^{-v_3} \right|_{\alpha v_1}^{\infty} \mathrm{d}v_1 \\
&= e^{w_1+w_3} \int_{v_1 \geq w_1} e^{-(1+\alpha)v_1} \mathrm{d}v_1 \\
&= e^{w_1+w_3} \left| -\frac{1}{(1+\alpha)} e^{-(1+\alpha)v_1} \right|_{w_1}^{\infty} \\
&= e^{w_1+w_3} \frac{1}{1+\alpha} e^{-(1+\alpha)w_1} \\
&= \frac{1}{1+\alpha} e^{-\alpha w_1 + w_3}
\end{aligned} \tag{20}$$

The proof of lemma 2 now consists of the chain of equivalences in (21). Step (21a) holds because the underlying form x_3 has only two candidates y_3 and z_3, whereby the probability mass of z_3 is equal to 1 minus the probability mass of y_3. Step

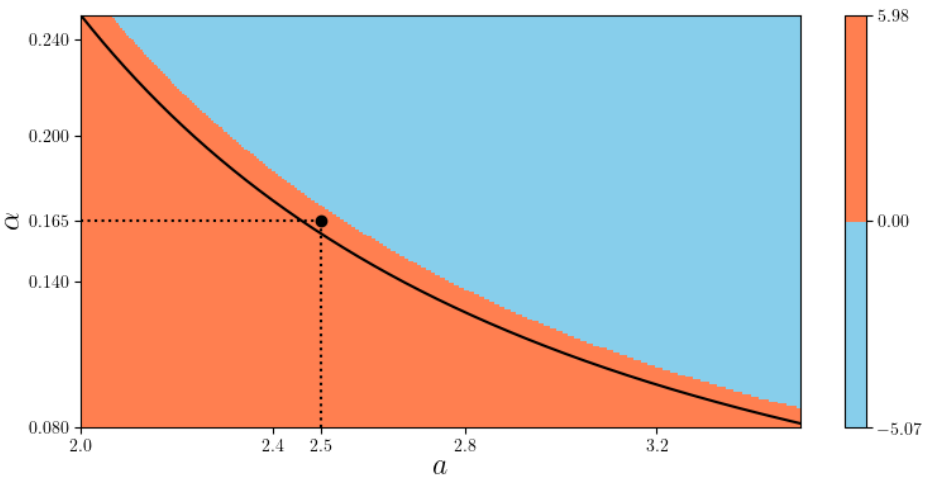


Figure 3

(21b) holds because the probability mass of the mapping $(\mathsf{y}_3 \,|\, \mathsf{x}_3)$ according to the SHG grammar corresponding to the weight vector $\boldsymbol{w}$ is the volume of the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ relative to the product of three exponential densities that start at the weights w_1, w_2, and w_3. Step (21c) holds because the definition (10) of the cone $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$ only looks at the first and third components. Step (21d) holds because of the computation in (20) above.

$$
\begin{aligned}
& G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{y}_3 \,|\, \mathsf{x}_3) > G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{z}_3 \,|\, \mathsf{x}_3) \iff \\
& \quad \overset{(a)}{\iff} \quad G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{y}_3 \,|\, \mathsf{x}_3) > 1 - G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{y}_3 \,|\, \mathsf{x}_3) \\
& \quad \iff \quad 2 G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{y}_3 \,|\, \mathsf{x}_3) > 1 \\
& \quad \overset{(b)}{\iff} \quad 2 \int_{\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)} p^{\text{exp}}_{\boldsymbol{w}}(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v} > 1 \\
& \quad \overset{(c)}{\iff} \quad 2 \int_{\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)} p^{\text{exp}}_{w_1}(v_1) \, p^{\text{exp}}_{w_3}(v_3) \, \mathrm{d}v_1 \, \mathrm{d}v_3 > 1 \\
& \quad \overset{(d)}{\iff} \quad 2 \frac{1}{1+\alpha} \exp(w_3 - \alpha w_1) > 1 \\
& \quad \iff \quad w_3 > \alpha w_1 + \log \frac{1+\alpha}{2}
\end{aligned}
\tag{21}
$$

C Proof of lemma 3

Step (22a) holds as steps (21a-c) above. Step (21b) can be established by reasoning as in (20).

$$
\begin{aligned}
& G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{y}_1 \,|\, \mathsf{x}_1) > G^{\text{SHG}}_{\boldsymbol{w}}(\mathsf{z}_1 \,|\, \mathsf{x}_1) \\
& \quad \overset{(a)}{\iff} \quad 2 \int_{\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)} p^{\text{exp}}_{w_1}(v_1) \, p^{\text{exp}}_{w_2}(v_2) \, \mathrm{d}v_1 \, \mathrm{d}v_2 \\
& \quad \overset{(b)}{\iff} \quad 2 \left(1 - \frac{a}{1+a} \exp - \frac{w_1 - a w_2}{a} \right) > 1 \\
& \quad \iff \quad w_1 > a w_2 + a \log \frac{2a}{1+a}
\end{aligned}
\tag{22}
$$

D Proof of lemma 4

The positions $w_1 = a w_2 + \epsilon B$ and $w_2 = a w_3 + \epsilon B$ satisfy the second and third inequalities in (15) as long as $\epsilon > 1$. Plugging the latter into the former yields $w_1 = a^2 w_3 + \epsilon B(a+1)$. Plugging the latter into the first inequality in (15) yields (23).

$$
\big(\alpha a^2 - 1\big) w_3 < A - \alpha \epsilon B (1 + a) \tag{23}
$$

The assumption $a^2 \alpha > 1$ means that the coefficient of w_3 on the left-hand side of (23) is strictly positive. Hence, (23) admits a non-negative solution $w_3 \geq 0$ provided $A - \alpha\epsilon(a+1)B > 0$. Equivalently, provided ϵ satisfies (24). And the latter in turn requires (16), because $\epsilon > 1$.

$$
1 < \epsilon < \frac{A}{\alpha(1+a)B} \tag{24}
$$

In conclusion, non-negative solutions $w_1, w_2, w_3 \geq 0$ of the inequalities (15) can be constructed as follows. First, I choose a value ϵ that satisfies (24), which exists because of (16). Then, I construct $w_1, w_2, w_3 \geq 0$ backward as in (25). As desired, w_3 is non-negative because the numerator is non-negative by (24) and the denominator is positive because $a^2 \alpha > 1$ by (11).

$$
\begin{aligned}
w_3 &= \frac{1}{2} \frac{A - \epsilon\alpha(a+1)B}{\alpha a^2 - 1} \\
w_2 &= a w_3 + \epsilon B \\
w_1 &= a w_2 + \epsilon B
\end{aligned}
\tag{25}
$$

E Computing the cones

The following reasoning shows that, when the constraints are defined as in (5), the cone $\mathcal{W}(\mathsf{x}_1, \mathsf{y}_1)$

can be described through the inequality $v_1 > av_2$ in (10) with $a = 2.5$.

$$\begin{aligned} & \boldsymbol{v} \in \mathcal{W}(\mathsf{x}_1, \mathsf{y}_1) \\ \iff & \sum_{k=1}^{3} C_k(\mathsf{x}_1, \mathsf{y}_1) v_k < \sum_{k=1}^{3} C_k(\mathsf{x}_1, \mathsf{z}_1) v_k \\ \iff & v_1 > 2.5 v_2 \end{aligned}$$

An analogous reasoning holds for $\mathcal{W}(\mathsf{x}_2, \mathsf{y}_2)$ and $\mathcal{W}(\mathsf{x}_3, \mathsf{y}_3)$.

F Computing the weights

When a and α are chosen as in (19), the coefficients A and B defined as in (17) become $A = 0.540426093542$ and $B = 0.891687359847$. And condition (24) on ϵ becomes (1).

$$(1) \quad 1 < \epsilon < \frac{A}{\alpha(1+a)B} = 1.04947406627$$

Thus, I can choose for instance $\epsilon = 1.03$. The weights in (6) are obtained from (25) with $a = 2.5$, $\alpha = 0.165$, and $\epsilon = 1.03$. These weights thus satisfy the three inequalities (12), (13), and (14).

References

Arto Anttila, Scott Borgeson, and Giorgio Magri. 2019. Equiprobable mappings in weighted constraint grammars. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 125–134. Association for Computational Linguistics.

Arto Anttila and Giorgio Magri. 2018. Does MaxEnt overgenerate? Implicational universals in Maximum Entropy grammar. In *AMP 2017: Proceedings of the 2017 Annual Meeting on Phonology*, Washington, DC. Linguistic Society of America.

Paul Boersma. 1997. How we learn variation, optionality and probability. In *Proceedings of the Institute of Phonetic Sciences (IFA) 21*, pages 43–58, University of Amsterdam. Institute of Phonetic Sciences.

Paul Boersma. 1998. *Functional Phonology*. Ph.D. thesis, University of Amsterdam, The Netherlands. The Hague: Holland Academic Graphics.

Paul Boersma and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox Press, London.

Canaan Breiss and Adam Albright. 2022. Cumulative markedness effects and (non-)linearity in phonotactics. *Glossa: a journal of general linguistics*, 7:1–32.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the Stockholm Workshop on Variation Within Optimality Theory*, pages 111–120, Stockholm University.

Bruce Hayes and Aaron Kaplan. 2023. Zero-weighted constraints in Noisy Harmonic Grammar. *Linguistic Inquiry*, pages 1–14.

Bruce Hayes and Colin Wilson. 2008. A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440.

Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. 1990a. Harmonic Grammar – a formal multi-level connectionist theory of linguistic well-formedness: an application. In *Proceedings of the 12th annual conference of the Cognitive Science Society*, pages 884–891, Hillsdale, NJ. Lawrence Erlbaum Associates.

Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. 1990b. Harmonic Grammar – a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. In *Proceedings of the 12th annual conference of the Cognitive Science Society*, pages 388–395, Hillsdale, NJ. Lawrence Erlbaum.

Giorgio Magri and Arto Anttila. 2023. Paradoxes of MaxEnt markedness. In *AMP 2022: Supplemental Proceedings of the 2022 Annual Meeting on Phonology*. Linguistic Society of America.

Joe Pater. 2009. Weighted constraints in generative linguistics. *Cognitive Science*, 33:999–1035.

Alan Prince and Paul Smolensky. 1993/2004. *Optimality Theory: constraint interaction in generative grammar*. Blackwell, Oxford.

Brian W. Smith and Joe Pater. 2020. French schwa and gradient cumulativity. *Glossa: a journal of general linguistics*, 5:1–33.

Kie Zuraw and Bruce Hayes. 2017. Intersecting constraint families: an argument for Harmonic Grammar. *Language*, 93.3:497–546.