Varieties of Noisy Harmonic Grammar

Handout for talk at the Annual Meeting in Phonology (AMP)
University of Southern California, Los Angeles

BACKGROUND

1. Stochastic constraint-based grammar frameworks in modern linguistics
   - These generally originate with Optimality Theory (Prince and Smolensky 1993) and represent an effort to **stochasticize** it:
     - Generate not a single winner but rather a **probability distribution** over GEN.
   - The earliest frameworks of this sort were Partial Ordering OT (Anttila 1997) and Stochastic OT (Boersma 1997), but now there are a fair number of other approaches (below).
   - The pursuit of such frameworks is motivated by several considerations:

2. Gradience in language is pervasive (Bod et al. 2003)
   - Free variation outputs
   - Gradient well-formedness intuitions
   - Quantitative phonological patterns in the lexicon — often captured quantitatively by language learners and experimentally detectible (“Law of Frequency Matching”; Hayes/Zuraw/Siptár/Londe 2009, with lit. review)

3. Aside: is belief in non-gradience the consequence of traditional methodologies?
   - Often, fieldworker and theorist never meet.
   - The latter takes the former’s best-estimate conjecture as the analytical target.
   - Different results are obtained if the theorist is able to work with corpora (text and dictionary) or do experiments.

4. Stochastic grammar: analytic research is paired with learnability research
   - I.e. efforts at computational modeling of human language learning.

---

1 This talk develops material from a course at UCLA I co-taught with Kie Zuraw, to whom many thanks for help on this project. Thanks also to the members of the UCLA Phonology Seminar for listening to a rehearsal version of this talk and suggesting improvements.
• Typically each stochastic framework has coupled with it a computational procedure for learning its grammars.
• Stochastic frameworks are well adapted for this: isolated exceptions and speech errors don’t faze them.

SOME CONSTRAINT-BASED STOCHASTIC GRAMMAR FRAMEWORKS

5. I will focus on a subset of them here

• I focus on the frameworks that performed well in the testing carried out by Zuraw and Hayes.
  ➢ Kie Zuraw and Bruce Hayes (in press, Lg.), “Intersecting constraint families: an argument for Harmonic Grammar”

• Ability to predict the result when a single candidate competition responds to two independent constraint families strongly distinguishes the theories.
• The winners are all stochastic forms of …

6. Harmonic Grammar

• Legendre et al. (1990), Legendre et al. (2006), Potts et al. (2010), Pater (2016)
• Uses the same GEN-cum-EVAL architecture as Optimality Theory (Prince and Smolensky 1993).
• Constraints are not ranked but have numerical weights.
• For each candidate’s row in the tableau: multiply all violation counts by corresponding constraint weights and add up the total across constraints = Harmony, a kind of penalty.\(^2\)
• Winning candidate is the least penalized one.

<table>
<thead>
<tr>
<th>/Input/</th>
<th>CONSTRAINT1</th>
<th>CONSTRAINT 2</th>
<th>Harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>weights:</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Candidate 1</td>
<td>●*●</td>
<td>●*●</td>
<td>2 + 2 = 4</td>
</tr>
<tr>
<td>Candidate 2</td>
<td><em>●</em></td>
<td><em>●</em></td>
<td>2 + 2 = 4</td>
</tr>
</tbody>
</table>

7. Stochasticized versions of Harmonic grammar

• Two main approaches have been taken in “stochasticizing” this framework: NHG, and maxent.

---

\(^2\) Different scholars deploy minus signs differently, so for some Harmony is a negative quantity. The nomenclature/choice of signs followed here is taken from Wilson (2006).
8. Noisy Harmonic Grammar

- Ref.: Boersma and Pater (2008/2016)
- NHG **stochasticizes** Harmonic Grammar:
  - at each “evaluation time,” constraint weights are nudged upward or downward by a random amount drawn from a Gaussian distribution (bell curve) — the noise.\(^3\)
  - other than this preliminary step, selection of winner works the same as in non-stochastic Harmonic Grammar

9. Classical Noisy Harmonic Grammar as a procedure, shown as a tableau

- We add in the noise factors at the very start — to the constraint weights.
- Calculating, we get noisy harmonies.

<table>
<thead>
<tr>
<th>/Input/</th>
<th>CONSTRAINT1</th>
<th>CONSTRAINT 2</th>
<th>Harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight:</td>
<td>2 + N(_1)</td>
<td>1 + N(_2)</td>
<td></td>
</tr>
<tr>
<td>꙲ Candidate 1</td>
<td>2 \times (1 + N(_2))</td>
<td>2 \times (1 + N(_2))</td>
<td></td>
</tr>
<tr>
<td>Candidate 2</td>
<td>1 \times (2 + N(_1))</td>
<td>1 \times (1 + N(_2))</td>
<td>3 + N(_1) + N(_2)</td>
</tr>
</tbody>
</table>

- By iterating the noise-addition process, output probabilities of Candidate 1 and Candidate 2 can be estimated.

10. Varieties of Noisy Harmonic Grammar

- what I just described will be “Classical NHG”
- I will explore other variants, created by altering the procedure slightly:
  - Cell-specific noise
  - Addition of noise post violation-multiplication.\(^4\)

11. NHG with cell-specific noise (Goldrick and Daland 2009)

- We put in a **fresh noise value for every cell** (cells with no violations can safely be skipped).

<table>
<thead>
<tr>
<th>/Input/</th>
<th>CONSTRAINT1</th>
<th>CONSTRAINT 2</th>
<th>Harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight:</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>꙲ Candidate 1</td>
<td>2 \times (1 + N(_2))</td>
<td>2 \times (1 + N(_2))</td>
<td></td>
</tr>
<tr>
<td>Candidate 2</td>
<td>1 \times (2 + N(_1))</td>
<td>1 \times (1 + N(_3))</td>
<td>4 + N(_1) + N(_3)</td>
</tr>
</tbody>
</table>

- The difference is one of **scope**: at what level of calculation is the noise added in?
- We can think of (11) as displaying **cell-granularity** and (9) **constraint-granularity**.

---

\(^3\) This distribution is given an arbitrary standard deviation, which is the same for all constraints.

\(^4\) There are others; see Boersma and Pater (2016) for the option of letting weights go negative.
12. A completely different (non-NHG) approach to stochasticizing Harmonic Grammar: maxent grammars

- Again, carries forward the GEN-cum-EVAL architecture of OT.
- Again, carries forward Harmony, as calculated above, as the key basis of the theory.
- Intellectual ancestry: 19th century physics; 1980’s cognitive science (Smolensky 1986)
- A simple procedure converts Harmony to output probabilities:

\[ e^{\text{Harmony}(x)} \]

\[ Z = \sum e^{\text{Harmony}(j)} \]

\[ \text{probability:} \quad \frac{e^{\text{Harmony}(x)}}{Z} \]

- Intuitively:
  - Larger weights will have a greater role in lowering probability of violators.
  - Multiple violations will have more effect in lowering probability of violating candidate.

13. Where we are going

- We now have three stochastic theories in hand (2 NHGs and 1 maxent); more to come.
- Goal: locate cases in which these variant theories behave differently, in the hopes of connecting with empirical data that would bear on these differences.

HARMONIC BOUNDING

14. A very simple case of harmonic bounding

<table>
<thead>
<tr>
<th>/Input/</th>
<th>CONSTRAINT 1</th>
<th>CONSTRAINT 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate 1</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Candidate 2</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

- Candidate 2 has a proper superset of the violations of Candidate 1; i.e. Candidate 1 harmonically bounds Candidate 2.

15. Harmonic bounding in Classical OT

- Taught to beginners everywhere: harmonically bounded candidates never win under any ranking.

---

5 This clear and useful term from Colin Wilson is, alas, also a joke (eHarmony is a dating site on the web).
16. Harmonic bounding in Classical Noisy Harmonic Grammar

- Likewise, harmonically bounded candidates never win (frequency derived = zero).
- Why? In, e.g., (14) Candidate 2 must have a higher Harmony penalty than Candidate 1 on every evaluation time, and under any weighting, since the only cases where violations don’t match are to the disadvantage of Candidate 2.
  ➢ Exception: if weight of Constraint 1 is zero you get a 50/50 tie.

17. Maxent

- Harmonically bounded candidates can win in maxent (Jesney 2007)
- Maxent imposes a stochastic version of Harmonic Bounding:
  ➢ “A harmonically bounded candidate can never receive a higher probability than the candidate that bounds it.”

18. Relating constraint weights to probability of the harmonic-bounder candidate in maxent

- Easy to work out the math on a spreadsheet.
- For (14), weight of Constraint 2 does not matter at all — it cancels out.
- If the weight of Constraint 1 is zero, then candidates are 50/50.
- As the weight of Constraint 1 approaches infinity, probability of Candidate 1 asymptotes at 1.

19. Cell-granularity NHG

- Like maxent: harmonically bounded candidates are not shut out completely (Goldrick and Daland 2009)
- Because noise is computed separately for each candidate, in (14) Candidate 2 can sometimes win. These conditions must be met:
  ➢ By chance, Candidate 1 is penalized on CONSTRAINT 2 more than Candidate 2 is, for this particular evaluation time.
  ➢ And, the difference must exceed the weight of (randomly perturbed) CONSTRAINT 1.
• Working out the math, this too yields a rising curve asymptoting at 1.
• Maxent curve is given in gray for comparison. They are *almost* identical, when scaled appropriately.

![Graph showing probability vs. weight of constraint]

• Again, it is not possible for the more-penalized candidate to get a higher probability than its harmonic-bounder.
• So Maxent and Cell-granularity NHG stand out as frameworks that implement the probabilistic, rather than absolute interpretation of harmonic bounding.

**A SPECIAL CASE OF HARMONIC BOUNDING:**
**LOCAL OPTIONALITY**

**20. Premise**

• The same phonological process, guided by the same constraints, is applicable in more than one place in the input.

**21. Literature**

• There is a small pile of useful empirical cases to work with: French schwa deletion, Makonde vowel harmony, Bengali intonation phrasing, Pima reduplication ...
• AFAIK none of these has enough quantitative data to test the differences I will demonstrate.
22. I’ll use a schematic example: intervocalic voicing of /p/ in long strings

<table>
<thead>
<tr>
<th>/apapapapa/</th>
<th>IDENT(voice)</th>
<th>*VpV</th>
</tr>
</thead>
<tbody>
<tr>
<td>[apapapapa]</td>
<td>***</td>
<td>****</td>
</tr>
<tr>
<td>[apapapaba] et al.</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>[apapababa] et al.</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>[apabababa] et al.</td>
<td>***</td>
<td>*</td>
</tr>
<tr>
<td>[ababababa]</td>
<td>****</td>
<td></td>
</tr>
</tbody>
</table>

- We see **double pyramids** of asterisks, since reducing Markedness violates Faithfulness.
- “Double pyramids” are characteristic, since multiple application will generally involve tradeoff of Markedness and Faithfulness violations.

23. Behavior of classical NHG

- Derives *only the extreme candidates*.
- Why? Harmonic bounding, as previous scholars have noted.

24. Multiple-violation patterns: what of the theories with probabilistic harmonic bounding?

- Maxent and Cell-granularity NHG *can* assign probability to the medial candidates.
- But let’s take this further: what sort of *probability distributions* to the theories characteristically assign?

25. An abstract example to work with

- We scale up (22) to *six* /p/’s, with underlying /apapapapapapa/.
- Assume 50% probability of intervocalic voicing in all positions, with independent outcome for each locus.⁶

---

⁶ I am aware that a real-life counterpart of this language, Warao (Osborn 1966), is claimed to have obligatory across-the-board-or-none application. On the slimness of the Warao data see Riggle and Wilson (2005).
Here is a tableau for simulation/study, with the 64 equiprobable candidates.

<table>
<thead>
<tr>
<th>Input</th>
<th>Candidate</th>
<th>Freq.</th>
<th>*VpV</th>
<th>IDENT(voice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/apapapapapa/</td>
<td>apapapapapapa</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>apapapapapaba</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>apapapapabapa</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>apapapabapapa</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>apapabapapapa</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>apabapapapapa</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>abapapapapapa</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>apapapapababa</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>apapabapababa</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>(48 not listed)</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td></td>
<td>ababababababa</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>ababababababa</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

26. **Maxent**

- Under suitable training conditions,\(^7\) we get the most minimal analysis conceivable.
  - Both constraints are weighted **zero**.
  - Equiprobability is predicted — a perfect fit.
- Indeed, this is why the theory is called maximum entropy!
  - = Minimal commitments in the absence of any sort of pattern in the training data.

27. **Cell-granular NHG**

- I have implemented Cell-granular NHG on a private copy of my “OTSoft” software,\(^8\) with the Boersma-Pater learning algorithm.
- Curiously, probability is allocated to all candidates, but with *priority to the extremes*, and lowest probabilities to the midpoints.
  - The candidates are sorted here by number of /p/ → [b] changes; i.e. in order of descending Faithfulness.

---

\(^7\) Specifically, the general penalty for constraint weights (Gaussian prior) often employed for maxent.

\(^8\) www.linguistics.ucla.edu/people/hayes/otsoft/
28. Naming this pattern

- **Upsilonism**: the property of learning a U-shaped probability distribution when trained on flat data.
  - This is a property of cell-granular NHG (and no other theory I know of)

29. An intuitive explanation of upsilonism in cell-granular NHG

- Think of just the candidate [apapapapapapa].
- Suppose that the roll of the noise-dice is lucky for this candidate: the constraint that penalizes it, *VpV, happens to get a low weight.
- [apapapapapa] benefits *sixfold* from the lowness of this weight.
- [abababababa] likewise benefits sixfold when IDENT(voice) gets a low weight for this cell.
- “Medial” candidates benefit less (fivfold, fourfold, etc.) from their lucky moments — so they have trouble standing out; the more so the closer to the center they lie.

30. Flipping the terms of the deal: maxent deals with U-shaped training data

- Experiment: I trained a maxent grammar
  - same constraints, candidates, violations
  - frequencies were the very frequencies output by the upsilonic cell-granular NHG
- Result: flatline! — equal frequencies assigned across the board
- We might call this deviation-from-training-data **rectilinearism**.

31. A pattern that might someday lead to empirical testability

- Multiple-locus phonology
- Opposed constraints of Markedness and Faithfulness
• Training data is overall symmetrical between faithful and altered outputs.
• Maxent predicts rectilinearism; cell-granular NHG predicts upsilonism.

STILL MORE VARIETIES OF NHG:
POST-MULTIPLICATIVE VS. PRE-MULTIPLICATIVE NOISE ADDITION

32. The system of noise-addition for Classical NHG (repeated)

<table>
<thead>
<tr>
<th>Input</th>
<th>CONSTRAINT 1</th>
<th>CONSTRAINT 2</th>
<th>Harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight:</td>
<td>3 + N₁</td>
<td>1 + N₂</td>
<td></td>
</tr>
<tr>
<td>Candidate 1</td>
<td>2 × (1 + N₂)</td>
<td>2 + (2 × N₂)</td>
<td></td>
</tr>
<tr>
<td>Candidate 2</td>
<td>1 × (3 + N₁)</td>
<td>1 × (1 + N₂)</td>
<td>4 + N₁ + N₂</td>
</tr>
</tbody>
</table>

• Notice that N₂ got doubled in the final Harmony computation for Candidate 1.

33. Changing scope again: noise addition follows violation-multiplication

<table>
<thead>
<tr>
<th>Input</th>
<th>CONSTRAINT 1</th>
<th>CONSTRAINT 2</th>
<th>Harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic constraint weight:</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>select and remember a noise value:</td>
<td>N₁</td>
<td>N₂</td>
<td></td>
</tr>
<tr>
<td>Candidate 1</td>
<td>(2 × 1) + N₂</td>
<td>(2 × 1) + N₂</td>
<td>2 + N₂</td>
</tr>
<tr>
<td>Candidate 2</td>
<td>(I × 2) + N₁</td>
<td>(I × 1) + N₂</td>
<td>4 + N₁ + N₂</td>
</tr>
</tbody>
</table>

• We pick a noise value for each constraint.
• But add it in only after the violations have been multiplied by weights.
• The harmonies emerge as different: unmultiplied weights
  ➢ Difference here: Harmony(Candidate 1) is 2 + N₂ instead of 2 + 2N₂

34. Terminology

• Classical NHG has premultiplicative noise.
• (33) is postmultiplicative noise.

35. A scenario in which a difference can be observed

• Assume binary competitions in which two viable candidates, 1 and 2, compete for each input.
• Assume one constraint that always incurs one violation and penalizes Candidate 1.
• Assume an opposed scalar constraint, with a range of integer values, penalizing Candidate 2 depending on “how many” of something it has.

36. McPherson and Hayes’s (2016) example: Tommo So vowel harmony

• The single-violation constraint is IDENT(feature) (Faithfulness)
• The scalar constraint is \textsc{Agree(feature)} — for seven degrees of “morphological closeness”
  ➢ Formalizing the “levels” of classical Lexical Phonology (Kiparsky 1982)
  ➢ Inner levels give more violations of \textsc{Agree}.

37. Toy example for the comparisons to follow

• We’ll do a highly schematic tableau; real-life cases tend to be cut off owing to lack of the full range of input.

<table>
<thead>
<tr>
<th></th>
<th>VARIABLE CONSTRAINT</th>
<th>CONSTANT CONSTRAINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input 1</td>
<td>Cand1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cand2</td>
<td>0</td>
</tr>
<tr>
<td>Input 2</td>
<td>Cand1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cand2</td>
<td>1</td>
</tr>
<tr>
<td>Input 3</td>
<td>Cand1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cand2</td>
<td>2</td>
</tr>
<tr>
<td>Input 4</td>
<td>Cand1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cand2</td>
<td>3</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Input 20</td>
<td>Cand1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Cand2</td>
<td>19</td>
</tr>
</tbody>
</table>

38. The essential plot — maxent version

• How does the probability of Candidate 1 go down as we increase the number of violations of \textsc{variable constraint}?
• The following chart is for \( \text{weight}_{\text{variable}} = 1, \text{weight}_{\text{constant}} = 10 \)

• This is a \textbf{sigmoid} curve.
• Mathematically it is a version of the \textbf{logistic function}. 
- It is **symmetrical** about the point of 50% probability.
- It asymptotes at **one** and **zero**.
- It fits the Tommo So data pretty well; see McPherson/Hayes.

39. Same example, NHG with post-multiplicative noise

- The following chart is for $\text{weight}_{\text{variable}} = 2$, $\text{weight}_{\text{constant}} = 20$ (twice the maxent weights).

![Graph](image)

- This, too, is a sigmoid curve.
- Mathematically it is a version of the **cumulative normal distribution** — amazingly similar to the logistic function but of completely different mathematical origin.
- It too is symmetrical about the 50% point and asymptotes at one and zero.
- It also fits the Tommo So data pretty well.

40. Classical NHG (pre-multiplicative noise)

- Weights again 2 for constant constraint and 20 for variable constraint.
• This one is different:
  ➢ Not symmetrical.
  ➢ Asymptotes at one on the left, but not at zero on the right.
  ➢ The asymptote at the right turns out to be about 0.07.

41. Why these differences? I: maxent

• Maxent: symmetricalness about 50% point and zero/one asymptotes are direct consequences of the math.
  ➢ Supplemental Materials for McPherson and Hayes (2016) gives all the details.9

42. Why these differences? II: post-multiplicative NHG

• Imagine two Gaussians;
  ➢ the one for the variable constraint “slides rightward” as the number of violations of the variable constraint goes up.
  ➢ They keep the same standard deviation, because noise is not multiplied.
  ➢ Probability of a less-likely outcome (“reversed” harmony) will asymptote to zero, as the two Gaussians separate:

43. Why these differences? III: Classical (pre-multiplicative) NHG

• Again, sliding Gaussians for probability distributions of weights.
• But the rightward-sliding Gaussian also expands its standard deviation as it slides.
• So some “left tail” that has some reasonable chance of overlapping with the constant-constraint Gaussian remains in place, so probability eventually settles on an above-zero minimum.10

---

9  www.linguistics.ucla.edu/people/hayes/papers/SupplementalFilesMcPhersonAndHayes2016.zip
10 For the math involved here consult http://mathworld.wolfram.com/NormalDifferenceDistribution.html.
44. Our tweaks are freely combinable

- It’s perfectly feasible to set the noise afresh in every tableau cell, and add it in after weights are multiplied by violations, thus:

<table>
<thead>
<tr>
<th>/Input/</th>
<th>CONSTRAINT1</th>
<th>CONSTRAINT 2</th>
<th>Harmony</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight:</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Candidate 1</td>
<td>(2 × 1) + N₂</td>
<td>2 + N₂</td>
<td></td>
</tr>
<tr>
<td>Candidate 2</td>
<td>(1 × 3) + N₁</td>
<td>(1 × 1) + N₃</td>
<td>4 + N₁ + N₃</td>
</tr>
</tbody>
</table>

- This creates yet another variety of Noisy Harmonic Grammar — post-multiplicative-noise, cell-granularity

45. The /apapapapapa/ language in post-multiplicative-noise, cell-granularity NHG

- Result:
  - Moving to post-multiplicative noise cancels upsilonism: output distribution is flat, just like in maxent.
  - This makes sense: the intuitive explanation for upsilonism in (29) depended on the multiplication of the noise.
- Rectilinearity:
  - Like maxent, post-multiplicative-noise, cell-granularity NHG has a rectilinearity bias: it outputs a flat distribution when trained on an upsilonic one.
46. Accidental twins?

- I have yet to find any qualitative differences between maxent and post-multiplicative noise, cell-granularity NHG; though their math is entirely different.\(^\text{11}\)
- As noted, there are differences but they rest on the subtle difference between logistic vs. normal-distribution sigmoids — far beyond the resolution of any sort of empirical work being done today.

47. Summary: classifying the differences

<table>
<thead>
<tr>
<th>Framework</th>
<th>Effects of harmonic bounding</th>
<th>Upsilonism/rectilinearism/ extremes only</th>
<th>Type of sigmoids derived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical NHG</td>
<td>absolute</td>
<td>extremes only</td>
<td>asymmetrical with non-zero asymptote</td>
</tr>
<tr>
<td>NHG with cell-specific noise</td>
<td>probabilistic</td>
<td>upsilonism</td>
<td>asymmetrical with non-zero asymptote</td>
</tr>
<tr>
<td>NHG with post-multiplicative noise addition</td>
<td>absolute</td>
<td>extremes only</td>
<td>symmetrical with 1/0 asymptotes</td>
</tr>
<tr>
<td>NHG with cell-specific noise and post-multiplicative weight addition</td>
<td>probabilistic</td>
<td>rectilinearism</td>
<td>symmetrical with 1/0 asymptotes</td>
</tr>
<tr>
<td>Maxent</td>
<td>probabilistic</td>
<td>rectilinearism</td>
<td>symmetrical with 1/0 asymptotes</td>
</tr>
</tbody>
</table>

THE EMPIRICAL SIDE

48. What I have something to say about ..

- I have no data to bear on upsilonism/rectilinearism.
- I will address only the issues of
  - statistical harmonic bounding
  - sigmoid shape

49. Inventory theory

- Linguistics is filled with cases where derivations are of dubious value.
- I.e. we simply want to indicate the legal members of an inventory (possible words, possible sentences, possible metrical lines of verse)
  - Cf. phonology, where the task of unifying the members of a paradigm give us reason to derive surface forms from underlying forms.

---
\(^{11}\) I just had time to check Jesney’s (2007) case (CCVC can lose its coda or simplify its onset independently). Again they behave alike.
• The Rich Base concept does permit “pseudo-derivations” to underlie inventories, with a literature addressing this (Keer and Baković 1997, Hayes 2004, Prince and Tesar 2014, etc.)
• An alternative is simply to develop a Faithfulness-free stochastic grammar that assigns a probability to every member of GEN.
• Such a grammar would assign, I suspect, a lower probability to [pra] than [pa];
  ➢ [pra] is harmonically bounded, and indeed should by any criterion receive a lower probability score than [pa].
  ➢ Pseudo-derivational theories face the awkward fact that [pra] is more marked than [pa] but not repaired (other than in rare speech errors) to [pa].
• All such work must assign some probability to harmonically bounded candidates, and thus works only with some of the frameworks above.

50. An apparent case of statistical harmonic bounding

• Hayes and Schuh (in progress) evaluate three of the principal “metra” of the Hausa *rajaz* quantitative meter.

<table>
<thead>
<tr>
<th></th>
<th>STRONG POSITIONS MUST INITIATE HEAVY SYLLABLE</th>
<th>HEAVY SYLLABLE MUST INITIATE STRONG POSITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>x x x x x x</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>x x x x x x</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>x x x x x x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x x x x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x x x x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• A near-perfect pattern: no matter what the actual frequencies, the relative frequencies respect harmonic bounding relations for various poets/poems/stanza positions.

![Graph showing characteristic shape of sigmoids in grammar.]

• Why? because \( \sim \sim \sim \) bounds \( \sim \sim \) and \( \sim \sim \) bounds \( \sim \sim \sim \); and in the frameworks adopted here, subsets of violation implies *invariant differences of probability*; rather than outright absence of the candidate with superset violations.

51. The Goldrick/Daland theory of speech errors

• If Goldrick/Daland are right that a “misweighted” grammar occasionally outputs harmonically-bounded candidates.

52. Multiple locus cases

• Already discussed.
• To be sure, there are other remedies available. (Riggle-Wilson; Kaplan; Kimper)
• 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 (how probability distributions are computed)

**CHARACTERISTIC SHAPE OF SIGMOIDS IN GRAMMAR**

53. McPherson and Hayes’s (2016) result

• We tried to fit the Tommo So data using classical NHG (asymmetrical sigmoids) and got a somewhat inferior fit to the data.
• Reason: our empirical sigmoids were quite symmetrical.

54. The world in general

• Zuraw and Hayes (in press) looked at a fair number of empirical sigmoids in phonology.
• We repeatedly found symmetrical sigmoid curves asymptoting — if the empirically observed range permitted it — at zero and one.
• See Zuraw and Hayes (in press) for a brief survey of such sigmoids elsewhere in language.
  ➢ speech perception
  ➢ syntactic change (work of Kroch and colleagues)\(^{12}\)
  ➢ possibly even cognition in general

CONCLUSION

55. The research agenda

• Earlier research on stochastic constraint theories emphasized:
  ➢ simple feasibility (ability to fit particular data sets; e.g. Anttila 1997; Boersma and Hayes 2001)
  ➢ effectiveness of the associated learnability theory (e.g. Pater 2008)

• But it’s also possible to treat such theories as **abstract characterizations of possible grammars**, which permit/favor only particular data patterns.
  ➢ Zuraw and Hayes (in press) — only some theories capture the parallel sigmoid curves that characterize “intersecting constraint families”
  ➢ Here: empirical picture less clear, but the theories are quite different in their predictions re. harmonic bounding, upsilonism/rectilinearism, sigmoid shape

• Empirical work can now keep its eyes open for data bearing on relatively abstract principles of grammatical organization.

\(^{12}\) Indeed, I suspect that serious study of synchronic syntactic patterns from a stochastic constraint-based viewpoint would also reveal abundant sigmoidal phenomena.
References


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


