# The analysis of gradience in phonology: what are the right tools?

Bruce Hayes

UCLA

Workshop on Gradience, Stanford University, July 7, 2007

# 1. Background

# *1.1 What sort of formal models should we consider for the analysis of gradience?*

- Some contenders not discussed here:
  - **analogical models** (Skousen 2002, Bailey and Hahn 2001, Daelemans et al. 2004)
  - **connectionist models** (Rumelhart and McClelland 1986 et seq.)
- Focus here: "Quantitatively augmented" generative models

## *1.2 Quantitatively augmented generative models*

- Rules and constraints of generative grammar cover the primary descriptive work, and are adapted to gradience by *embedding* them in a quantitative framework.
- Such frameworks are usually couched in the language of **probability**.
- I will address two such models:
  - **Stochastic Optimality Theory**
  - **Maximum Entropy** models

## *1.3 Gradient model and algorithmic learning*

- Gradient analysis is hard, and we may be able to do better with machine-learned grammars.
  - Implemented systems can comb through the data, fine-tuning the grammar with greater care than humans can.
- Grammars learned by algorithm address the long-standing goal of generative theorizing, namely to explain how acquisition is possible.

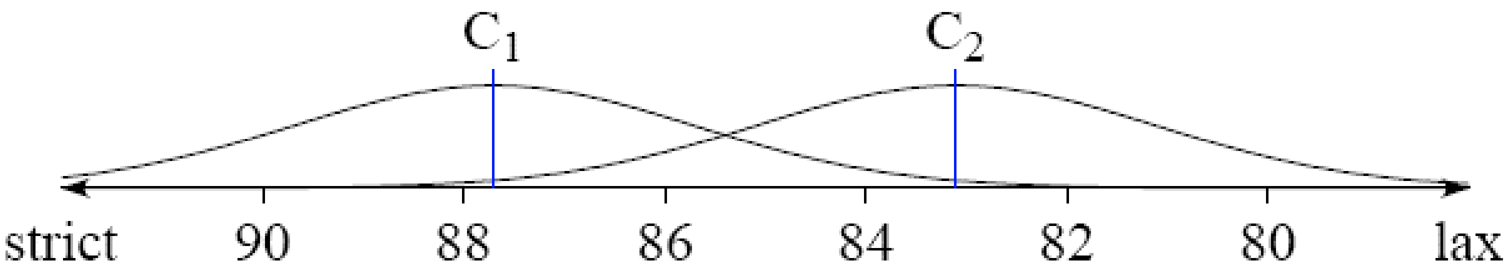
# 2. Stochastic Optimality Theory

- Refs.: Boersma (1997), Boersma and Hayes (2001)
- Basics:
  - ➢ Constraints are arranged in **ranking values** on a numerical scale, corresponding to their probability of being “ranked high.”
  - ➢ The ranking values define the means of Gaussian probability distributions, from which sampling takes place when the grammar is applied.

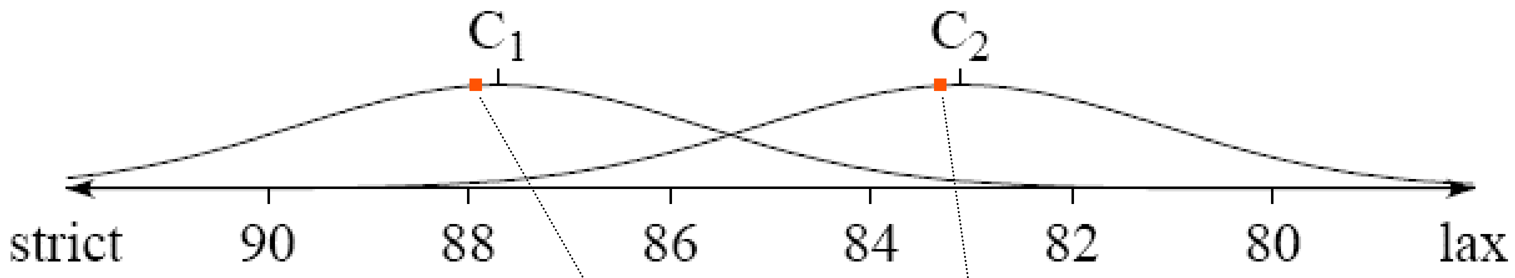
## *2.1 Example*

(taken from Boersma and Hayes 2001)

- Two constraints with distributions centered at the ranking values 87.7 and 83.1:



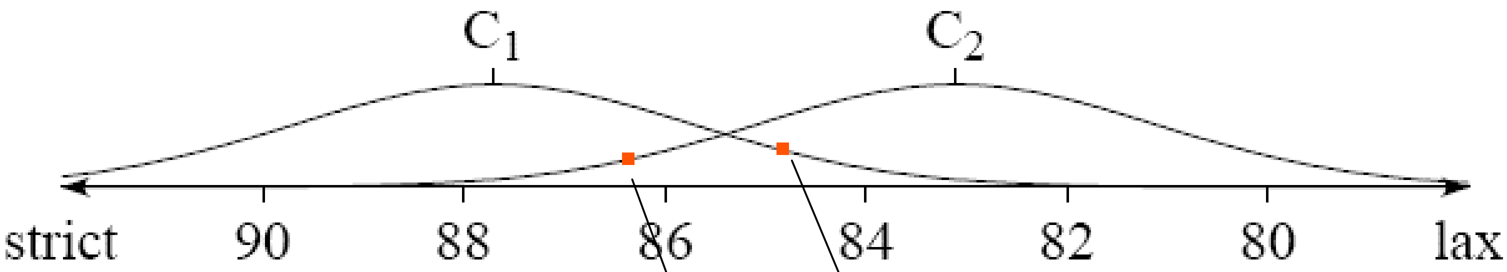
## 2.2 Sampling from the distributions and deriving a winner




- Obtain a selection point for each constraint by sampling.
- Sort constraint in descending order by selection point
- Find winner by normal methods of OT.

| /Input/ | C1 87.9 | C2 83.6 |
|---|---|---|
| ☞ Candidate1 | | * |
| Candidate2 | *! | |

## 2.3 A sample with the less-probable ranking and winner




| /Input/ | C2 85.1 | C1 83.4 |
|---|---|---|
| Candidate1 | *! | |
| ☞ Candidate2 | | * |

## *2.4 Long run behavior*

- This grammar generates Candidate1 94.8% of the time, Candidate2 5.2%.
- This is deducible analytically, or by simulation.

## *2.5 Learning Stochastic OT Grammars*

- **Starting point**
  - a constraint set
  - observed output forms with frequencies
  - suitable set of rival candidates for each input

## *2.6 Gradual Learning Algorithm (Boersma 1997)*

- Try the grammar on known input-output pairs. If it errs:
  - ➢ Incrementally raise the ranking values of all "winner-preferring" constraints
  - ➢ Incrementally lower the ranking values of all "loser-preferring" constraints.
- This has been shown in many cases to achieve good statistical matching to the learning data.

## *2.7 Other ranking algorithms for Stochastic OT*

- Maslova (to appear), Lin (2005), Wilson (2007)

# 3. Maximum Entropy grammars

- References: Eisner (2000), Johnson (2002), Goldwater and Johnson (2003), Hayes and Wilson (2007)
  - Closely related to Harmonic Grammar (Smolensky 1986, Smolensky and Legendre 2006) and more distantly to Linear Optimality Theory (Keller 2000, 2006).

## *3.1 Basics of Maximum Entropy grammars*

- Every constraint bears a *weight*, a nonnegative real number.
- The weight of a constraint specifies *a probability decrement* for candidates that violate it: "violating this constraint makes you *x* much less probable".

## *3.2 The math relating weights to output probabilities*

- **Step 1**: for each candidate *x* for an given input:
  - Compute its violations for each constraint $C_i$.
  - For each constraint $C_i$ multiply violations $C_i(x)$ times the weight of the constraint, $w_i$.
  - Sum the result over all constraints:

$$\Sigma_i \, w_i \, C_i(x)$$

- **Step 2**: take *e* to the negative power of the sum just calculated:

$$e^{-\Sigma_i w_i C_i(x)}$$

- **Step 3**: carry out similar sums for each candidate having the same input, and sum them. Call the result Z.

$$Z = \Sigma_y \left( e^{-\Sigma_i w_i C_i(y)} \right)$$

- **Step 4**: find the fraction of Z assigned to the candidate *x* under discussion:

$$\frac{\Sigma_x \exp(-\Sigma_i w_i C_i(x))}{Z}$$

This is the probability of candidate *x*.

# *3.3 Sample grammar*

| Text | Cand. | Target freq. | Predicted freq. | ME Score | C1 | C2 | C3 |
|---|---|---|---|---|---|---|---|
| | | | | | 32.7 | 11.0 | 0 |
| Input1 | 1-1 | 1 | 0.99999993 | 11.0 | | * | |
| | 1-2 | 0 | 0.00000007 | 32.7 | * | | |
| Input2 | 2-1 | 1 | 0.9999998 | 0 | | | * |
| | 2-2 | 0 | 0.0000002 | 11.0 | | * | |

- Weights were learned by algorithm; see below.
- Example calculation, first predicted score:

$$\frac{e^{-11.0}}{e^{-32.7} + e^{-11.0}} = 0.99999993$$, the practical equivalent of 1.

## *3.4 Learning*

- There are many ways to find the weights for a MaxEnt grammar.
- I cover here the method described and used in Hayes and Wilson (2007).
- This draws heavily on Della Pietra, Della Pietra, and Lafferty (1997).

# 3.5 Criterion

- Maximize the *probability of the observed data* P(*D*), given the constraint set (maximum likelihood estimation)
  - P(*D*) is the product of the probabilities of each observed datum, i.e. $\Pi_{x \in D}$ P($x$)

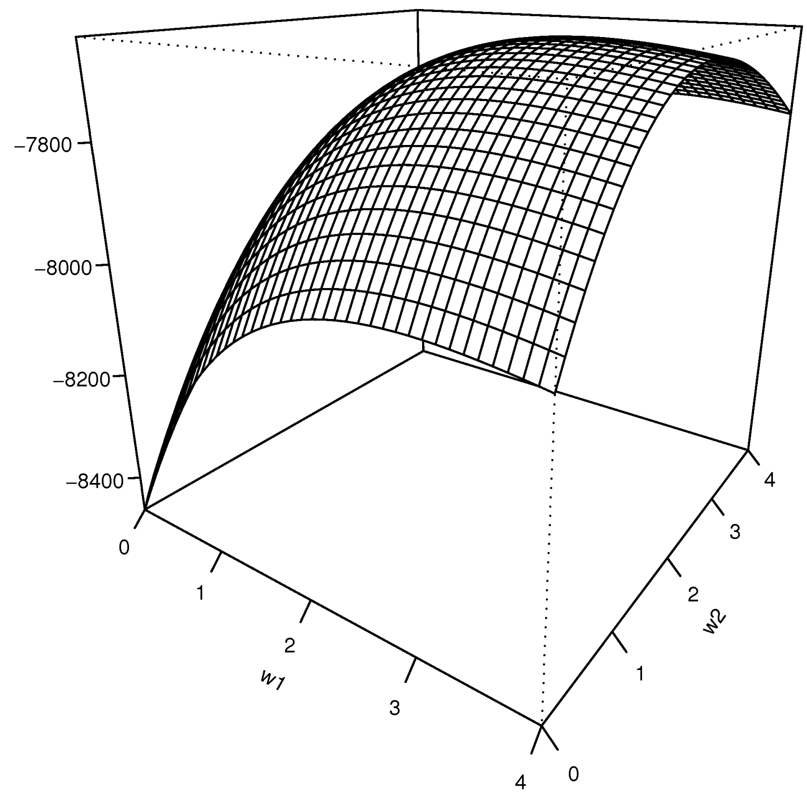
- This is a widely adopted criterion in learning theory.
- For an intuitive rationale, observe that it likewise *minimizes the probability of the unobserved data*, since probability sums to one for each input.

# *3.6 Method for maximizing probability of observed data*

- A hill-climbing search, conducted on an *n*-dimensional surface,
  - *n* = number of constraints.
  - “Altitude” is P(*D*).

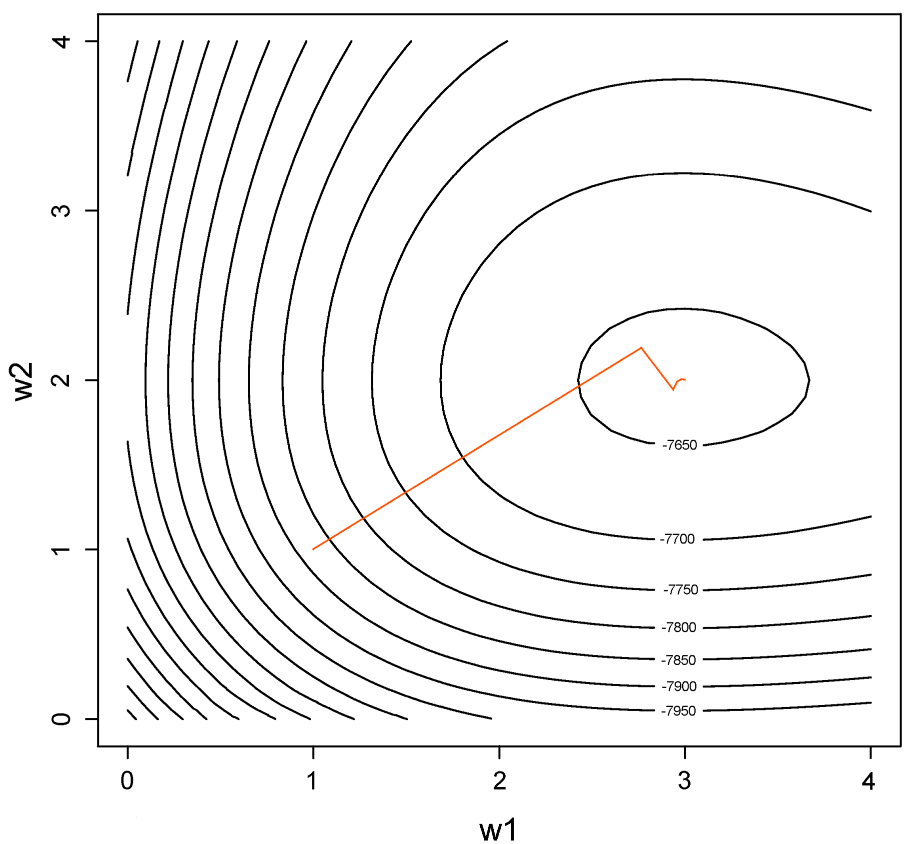
# 3.7 Example of hill-climbing

- Here, $n$ is 2, vertical axis depicts P($D$):



# *3.8 Climbing a hill stepwise: example*

- Contour map of same hill; top is at (3, 2).




| *Stage* | *Weight 1* | *Weight 2* |
|---|---|---|
| *1* | 1 | 1 |
| *2* | 2.77 | 2.19 |
| *3* | 2.94 | 1.94 |
| *4* | 2.96 | 1.99 |
| *5* | 2.98 | 2 |
| *6* | 2.99 | 2 |
| *7* | 3 | 2 |

## *3.9 How to climb*

- Climbing follows the **gradient**; i.e. vector of partial derivatives of (log) probability of observed data against individual weights ($\frac{\partial}{\partial w_i}$ log(P($D$))).

- A theorem due to Della Pietra et al. (1997) tells us how to compute the gradient: the component for each constraint is **O – E**, where
  - **O** = **observed** violation count in learning data
  - **E** = **expected** violation count (estimable from current guess for weights)

## *3.10 You won’t get lost…*

- Della Pietra et al. (1997) also demonstrate that the hill is **convex** (only one peak); hence no getting stuck in local maxima.

# *3.11 Convergence*

- Since the gradient is known, and the search space is convex, the weights found are guaranteed to be optimal; i.e. to maximize P(*D*).

## *3.12 Simulations reported here*

- Carried out with a software implementation of this algorithm created by Colin Wilson; public version in progress.

# 4. Some comparisons on general grounds

## *4.1 No Harmonic bounding in MaxEnt*

- In OT, any candidate that has a strict superset of a rival’s violations *never wins*.
- Not so in MaxEnt; see below.

# *4.2 Ganging*

- **Ganging effects**: when two constraints combine to overcome the effect of one competing constraint
  - Stochastic OT permits partial, **gradient** ganging effects (see Hayes and Londe 2006, 81, for a Hungarian example)
  - Maxent also permits outright **categorical** ganging.
  - For discussion of ganging, both empirical and theoretical, see Jäger and Rosenbach (2006), Keller (2000, 2006), McClelland and Van der Wyck (2006), Pater, Bhatt and Potts (2007).

## *4.3 A point of similarity*

- Every non-stochastic OT analysis has a MaxEnt equivalent (Johnson 2002, Prince 2002), but not vice versa (Smolensky and Legendre 2006, Pater, Bhatt and Potts, 2007), so the doubt is in the area of restrictiveness, not capacity.

## *4.4 Comparing learning algorithms*

- Unlike with MaxEnt learning, the support for GLA is purely "empirical": no proof has been found that it will find the best-fit ranking values for any data pattern.
- *Nor will there ever be*. Pater (in press) has constructed a clever counterexample:
  - an insidious pattern where many of the "winner preferrers" are, for other inputs, "loser preferrers", fatally confusing the GLA.
- The MaxEnt weighting algorithm given above easily learns Pater's data pattern, as I have checked.
- The unreliability of the GLA will be a factor in the discussion below.

## *4.5 MaxEnt weighting yields great precision*

- Boersma and Hayes's (2001) Ilokano simulation, redone in MaxEnt:

| *Output* | *Target frequency* | *GLA result* | *MaxEnt* |
|---|---|---|---|
| [tawʔen] | 1/2 | .489 | 0.50000006 |
| [taʔwen] | 1/2 | .511 | 0.49999990 |
| [buːbwaja] | 1/3 | .329 | 0.33333334 |
| [bwajbwaja] | 1/3 | .337 | 0.33333328 |
| [bubwaja] | 1/3 | .334 | 0.33333329 |

- This is of no importance for modeling experimental data (which has no such precision), but very helpful for diagnosing the adequacy of constraints.

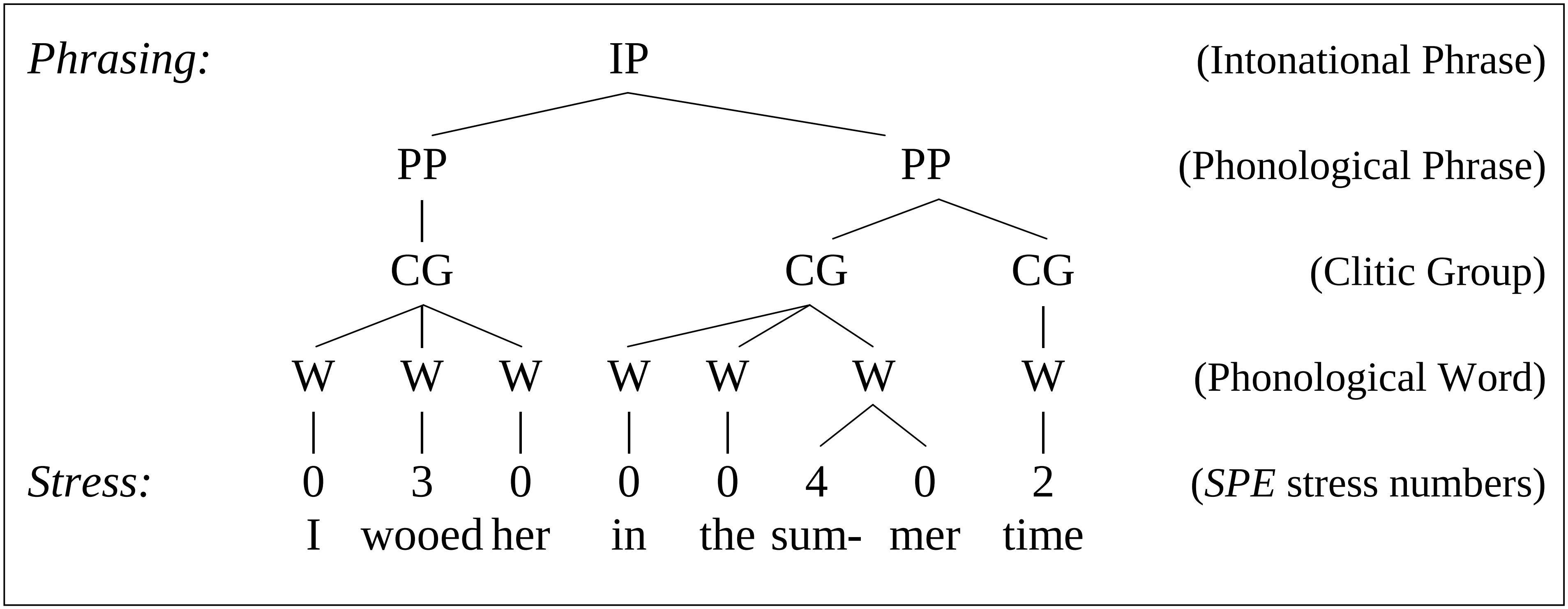
## *4.6 Remainder of this talk*

- Informal survey of my research life over the past two years—working with both GLA and MaxEnt to solve analytic problems.
- I don't yet know what the right tool for analysis of gradience is, yet, but hope that my experience is of interest.
- Cases:
    - Comparing the performance of stochastic OT/GLA with MaxEnt on a large simulation in **gradient metrics**.
    - Learning of **gradient phonotactics** (summarizing Hayes and Wilson 2007).

# 5. The textsetting problem

## 5.1 The textsetting problem

- Suppose we have a phonological representation, like this:



… and a rhythmic representation like this (Lerdahl and Jackendoff 1981)

```
    x       x       x       x
x   x   x   x   x   x   x   x
x x x x x x x x x x x x x x x x
```

- What should be the **temporal alignment of text to grid**?

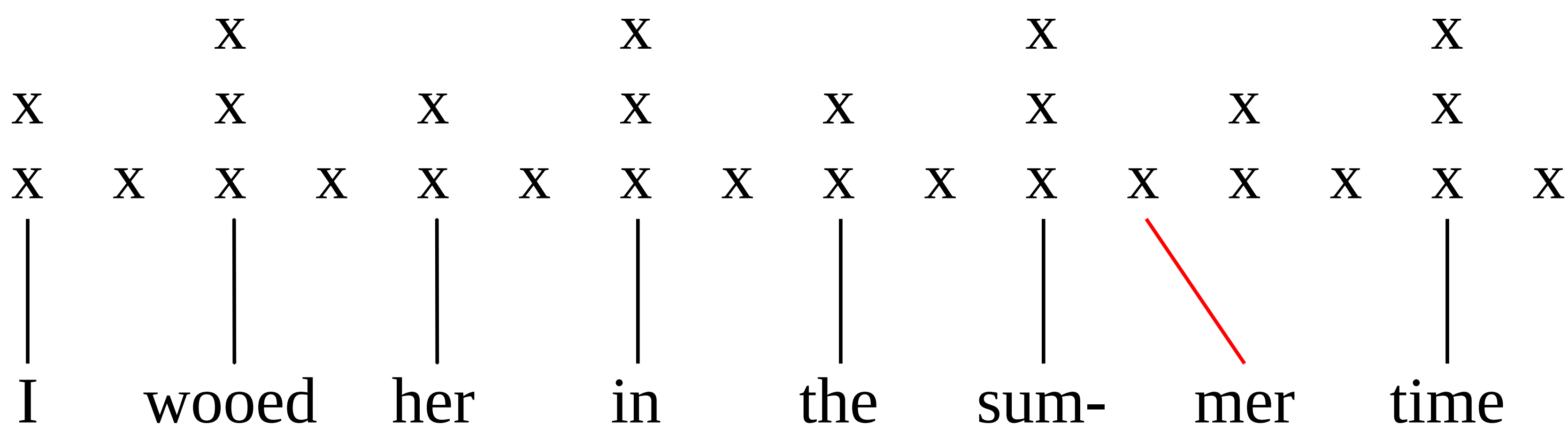
## *5.2 Possibilities I*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | x | | | | x | | | | x | | | | x | |
| x | | x | | x | | x | | x | | x | | x | | x | |
| x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| I | | wooed | | her | | in | | the | | sum- | | mer | | time | |

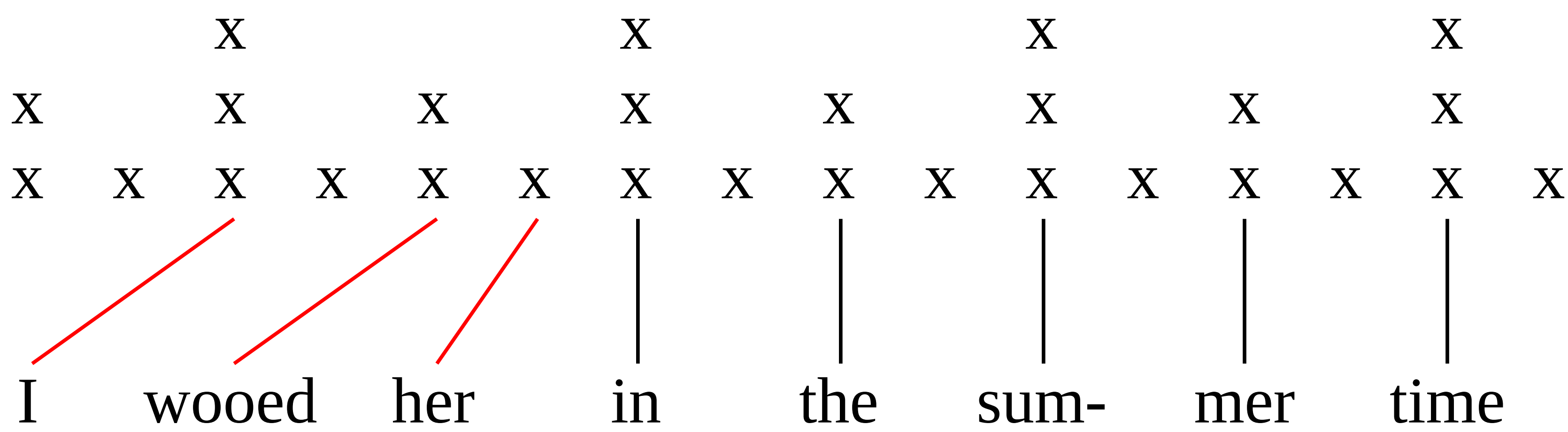
## *5.3 Possibilities II*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | x | | | | x | | | | x | | | | x | |
| x | | x | | x | | x | | x | | x | | x | | x | |
| x | x | x | x | x | x | x | x | x | x | x | x | x | x | x | x |
| I | | wooed | | her | | in | | the | | sum- | | mer | | time | |

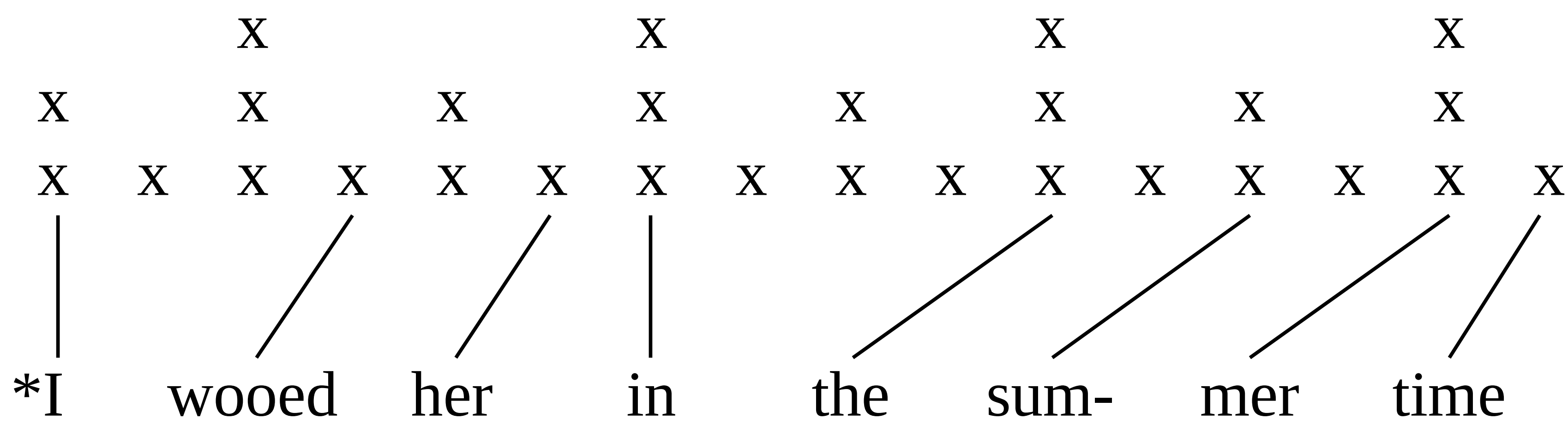
## 5.4 Possibilities III



## *5.5 Possibilities IV*



## *5.6 Not a possibility: one ill-formed textsetting*



## *5.7 The textsetting problem*

- People align texts to grids fluently—e.g. when they sing new verses to songs. How do they do it? How do they judge the well-formedness of settings?

- Textsetting is one problem in the field of **metrics**
  - see Halle and Keyser 1969, Kiparsky 1975, and much later work

- It is also a canonical area for **gradient analysis**:
  - usually multiple possibilities, which vary in preference
  - but also thousands of forms that must be fully excluded.

## *5.8 Previous work*

- A rule-based analysis:  Halle and Lerdahl (1993)
- An empirical study, with chanted settings elicited from nine native speaker consultants:  Hayes and Kaun (1996)
- A non-stochastic OT analysis, covering only "consensus" settings of the Hayes/Kaun corpus:  Hayes (in press)
- A preliminary stochastic OT grammar, learned with "easy", prefiltered data:  Hayes (2005)

## *5.9 My long-term plan for the study of textsetting*

- The theories of phonology and metrics will provide an appropriate **constraint set**.
- All the rest should follow from the choice of **framework**, particularly the learning algorithm.
- Exposure to different kinds of input data will result in differing textsetting **styles** or **dialects**, each the result of different stochastic rankings/weightings.

# *5.10 Learning simulations*

- Data taken from the Hayes/Kaun corpus (426 4-beat lines)
- Goal was to replicate the frequencies with which the consultants selected settings.
  - Hence values range from 0 (0/9) to 1 (9/9)
- Tools used:
  - Stochastic OT/GLA
  - MaxEnt

# *5.11 Constraints employed*

- These are the best constraint set I could devise for non-stochastic analysis (improving slightly on Hayes, in press)
- They serve three basic functions:

| *Match stress to rhythm* | *Regulate duration* | *Demarcate line division* |
|---|---|---|
| REGULATE SW<br>REGULATE SM<br>REGULATE MW<br>DON'T FILL W<br>MATCH LEXICAL STRESS<br>*MISMATCHED σ ˈσ]$_P$<br>FILL STRONG | RESOLUTION<br>STRONG IS LONG | *LAPSE<br>DON'T FILL 16<br>DON'T FILL 1 |

# *5.12 These constraints aren't bad*

- Applied to 364 lines with "consensus" votes, using nonstochastic OT to predict the most-selected scansion:
  - 267 successes, 90 misses, 7 ties
  - = about 3/4 correct

## *5.13 Training data*

- All 3592 textsettings (933 distinct) used by any consultant
- All 7,069 "contender" settings (Riggle 2004): those which, in OT, win or tie on at least one ranking.
- These two categories overlap heavily, but 408 attested settings (213 distinct) textsettings were not contenders.
- 40,000 other candidates, randomly selected from the ~4,000,000 logical possibilities.

# 5.14 Results (weights, ranking values)

| *Constraint* | *MaxEnt* | *GLA* |
|---|---|---|
| REGULATE SW | 14.02 | 106.0 |
| FILL STRONG | 12.28 | 108.0 |
| DON'T FILL16 | 4.78 | -1100.1 |
| *LAPSE | 4.07 | -1099.4 |
| DON'T FILL W | 3.81 | -3272.6 |
| DON'T FILL 1 | 2.08 | -3275.4 |
| MATCH LEXICAL STRESS | 1.97 | -1102.1 |
| RESOLUTION | 1.72 | -3278.7 |
| *MISMATCHED σ 'σ]$_P$ | 1.67 | -959.8 |
| REGULATE SM | 1.61 | -3274.8 |
| REGULATE MW | 0.94 | -3276.4 |
| STRONG IS LONG | 0.91 | -3328.9 |
| MATCH RISING LEX. STRESS | 0.14 | 49.8 |

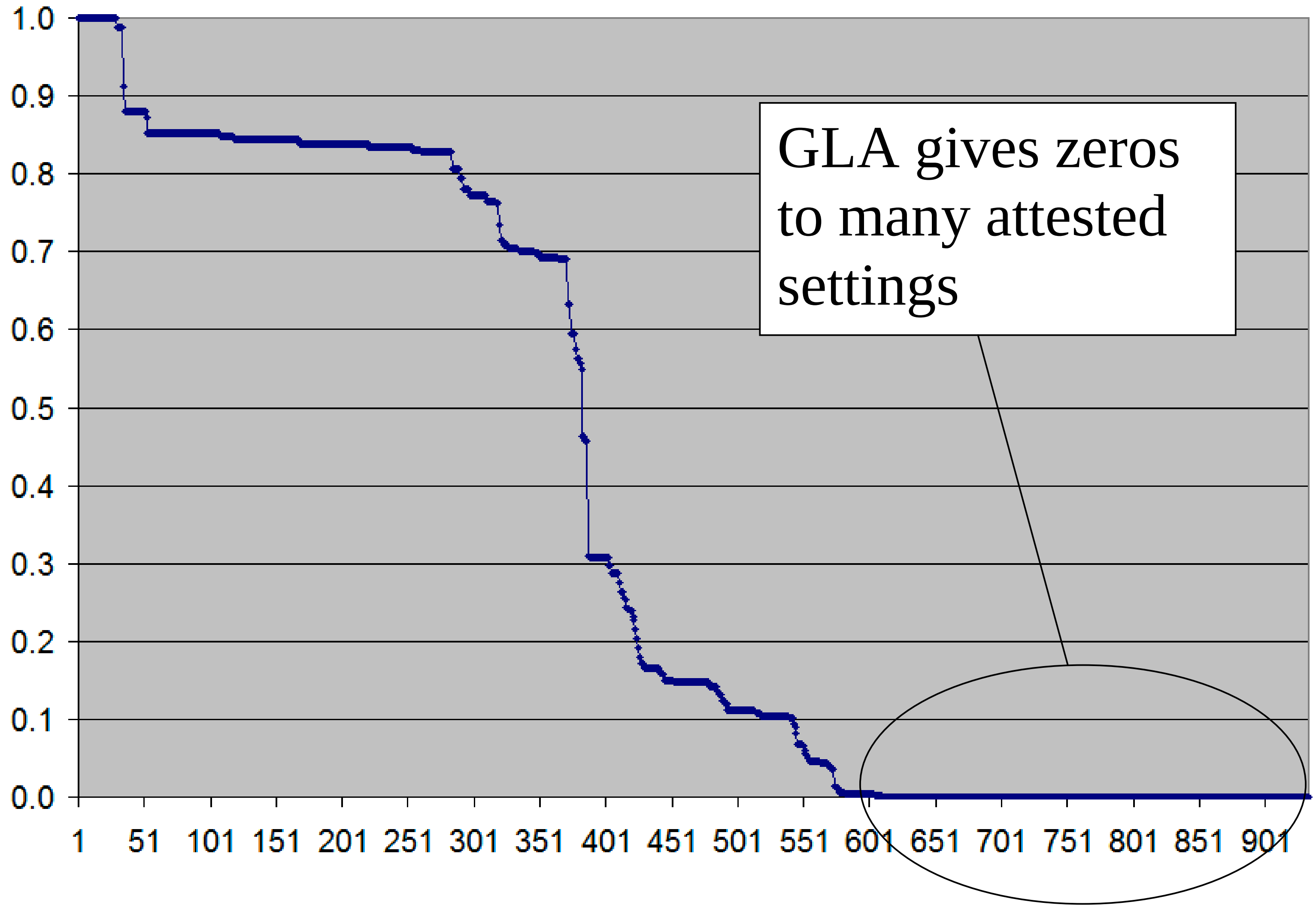
# *5.15 Comparison of grammar effectiveness*

- Correlation coefficients, all predicted frequencies vs. all observed, for two models. Not too bad, and also very similar!

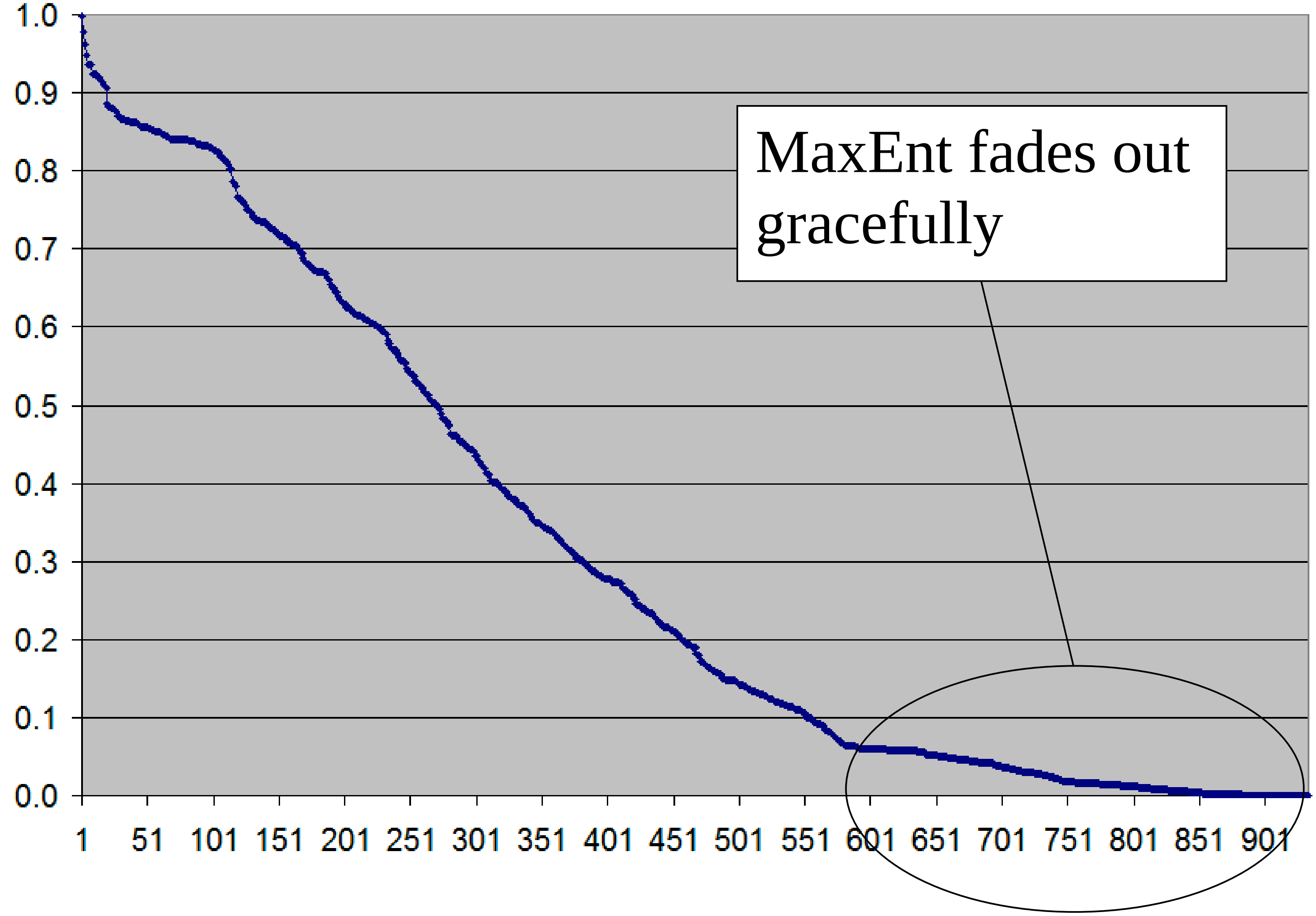
| | |
|---|---|
| MaxEnt grammar: | r = 0.843 |
| Stochastic OT grammar: | r = 0.841 |

- Nevertheless there is reason to think the maxent model is doing better:
  - the Stochastic OT/GLA model **assigns zero probability to too many settings**.

# *5.16 GLA predicted frequencies, sorted descending, for all settings volunteered by consultants*



# *5.17 MaxEnt predicted frequencies, sorted descending, for all settings volunteered*



# *5.18 Why does Stochastic OT/GLA gives zeros to so many attested settings?*

- Clearest answer: **ranking errors**
  - Some constraint pairs must be given **very close ranking values**, because they jointly determine common patterns of free variation.
  - But the GLA assigns them **very distant** values, corresponding to strict ranking.

# *5.19 Example of stochastic OT grammar failure*

- A common type of free variation (Hayes, 2005) requires free ranking of
  - STRONG IS LONG (give more time to strong beats)
    RESOLUTION (give little time to non-final stressed syllables)
- Tableau follows.

# *5.20 Tableau: a common kind of free variation*

```
     x       x          x        x
x    x   x   x    x     x    x   x
x  x x x x x x x  x  x  x x  x x x x
|  | |   |   | |  |     |    |   |
```

```
     x       x          x        x
x    x   x   x    x     x    x   x
x  x x x x x x x  x  x  x x  x x x x
|  | |   |   |    |  |  |    |   |
```

| | RESOLUTION | STRONG IS LONG |
|---|---|---|
| Such a pret-ty **sto**ry you soon shall hear | | * |
| Such a pret-ty **sto-** ry you soon shall hear | * | |

## *5.21 Ranking failure*

- The GLA placed these constraints very far apart:

| | |
|---|---|
| RESOLUTION | -3278.7 |
| STRONG IS LONG | -3328.9 |

(~ 50 units), and thus couldn't derive the second free variant.

- Since this variation pattern is common, this is a major source of the error of assigning too many zeros.

# *5.22 MaxEnt grammar does ok with these lines*

```
Line                                              Probability

     x       x          x        x
x    x   x   x    x     x    x   x
x  x x x x x x x  x  x  x x  x x x x
|  | |   |   | |  |     |    |   |
Such a pret-ty  story you   soon shall hear           .273
     x       x          x        x
x    x   x   x    x     x    x   x
x  x x x x x x x  x  x  x x  x x x x
|  | |   |   |    |  |  |    |   |
Such a pret-ty  sto-  ry you soon shall hear          .123
```

## *5.23 Another possible problem*

- The problem just noted was a problem with the GLA, trying to find the right ranking.
- But is there a Stochastic OT grammar that works *at all*??
- There would be none, if **harmonically bounded** candidates should be able to emerge with positive (though non-maximal) scores.
- Various constraint-based theories (Keller 2001, 2006; Coetzee 2006) do permit harmonically bounded candidates to win; MaxEnt is among them.

# *5.24 Harmonically bounded candidates in MaxEnt*

- MaxEnt allows harmonically bounded winners, though never with the highest frequency.
  - Simplest example, with just one constraint:

| Input | Cand. | Predicted Freq. | ME Score | C1 |
|---|---|---|---|---|
| | | | | 2.2 |
| Input | Cand1 | 0.9 | 0 | |
| | Cand2 | 0.1 | 2.2 | * |

## *5.25 Do harmonically bounded candidates win in textsetting?*

- About 11% of the settings volunteered by the consultants are not in the OT factorial typology of the constraint set.
- MaxEnt grammar gives most of these modest scores, averaging 0.04.
- Superficial implication: nonoptimal candidates can win, supporting MaxEnt.
- But this is *extremely tentative*—the problem could lie with the constraint set.

## *5.26 Summarizing*

- So far, MaxEnt is emerging as the better tool for my analytic purpose, due to:
  - Greater accuracy
  - Perhaps, its indulgence of harmonically bounded candidates

- But before drawing any conclusions we should try to learn more by
  - trying other stochastic ranking algorithms
  - exploring more possibilities for the constraint set.

# 6. Phonotactic learning with MaxEnt

## *6.1 My project with Colin Wilson*

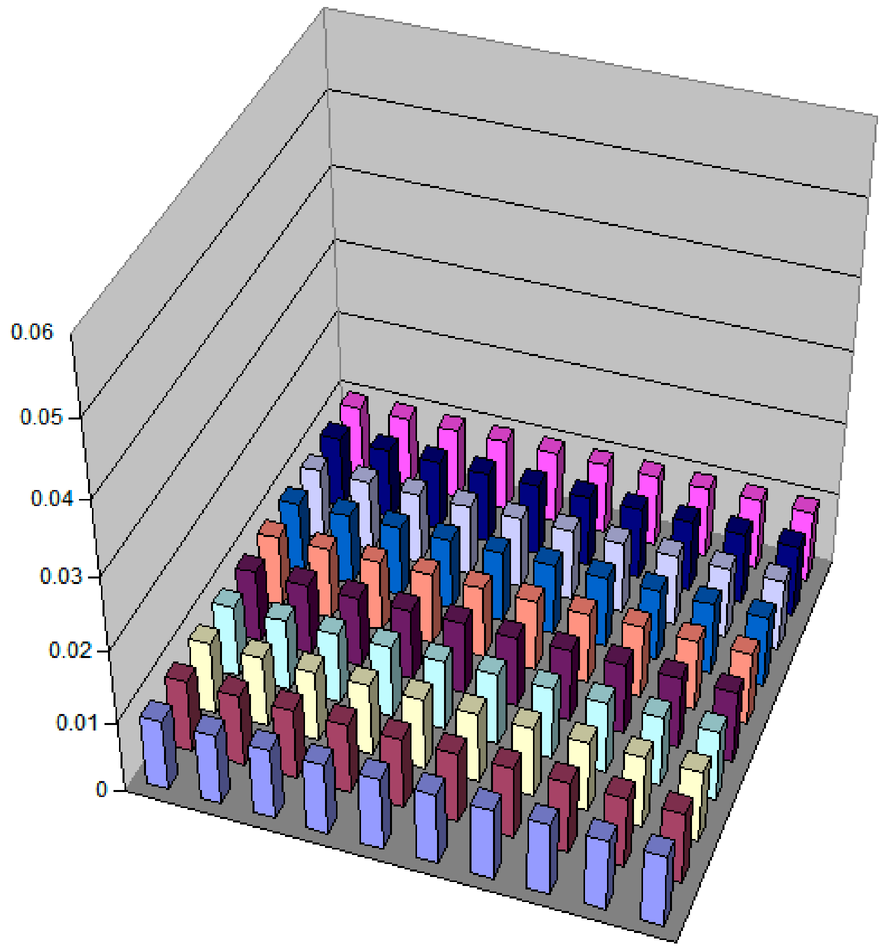
- We seek to produce an automated system that, examining representative phonological forms from languages will
  - learn a set of phonotactic constraints
  - weight them under the principles of MaxEnt
  - make accurate gradient predictions about the phonotactic well-formedness of any novel form

# 6.2 *Two kinds of probability distribution in phonology*

- Assume an input, and find the probability of possible corresponding outputs.
- What we do: assign probability **to all forms**.
  - Any one form will have an ultra-low probability, but the differences that exist among the ultra-low can be large and meaningful.
  - The problem of $\infty$ (unbounded string lengths) can be dealt with, for example by limiting the strings to (roughly) the length of those found in the learning data.

# *6.3 Graphic Illustration*

- Imagine the space of conceivable forms (here, just 100) to have equal *a priori* probability:



- A set of weighted phonological constraints penalizes various subsets. Here are four (schematic) ones, with their weights:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | * | * | * | * | * | * | * | * | * | * |
| 2 | * | * | * | * | * | * | * | * | * | * |
| 3 | * | * | * | * | * | * | * | * | * | * |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |
| 8 | | | | | | | | | | |
| 9 | | | | | | | | | | |
| 10 | | | | | | | | | | |

Weight: 3

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | * | * | * | * | | | | | | |
| 2 | * | * | * | * | | | | | | |
| 3 | * | * | * | * | | | | | | |
| 4 | * | * | * | * | | | | | | |
| 5 | * | * | * | * | | | | | | |
| 6 | * | * | * | * | | | | | | |
| 7 | * | * | * | * | | | | | | |
| 8 | * | * | * | * | | | | | | |
| 9 | * | * | * | * | | | | | | |
| 10 | * | * | * | * | | | | | | |

Weight: 4

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | * | * | * | * |
| 2 | | | | | | | * | * | * | * |
| 3 | | | | | | | * | * | * | * |
| 4 | | | | | | | * | * | * | * |
| 5 | | | | | | | * | * | * | * |
| 6 | | | | | | | * | * | * | * |
| 7 | | | | | | | * | * | * | * |
| 8 | | | | | | | * | * | * | * |
| 9 | | | | | | | * | * | * | * |
| 10 | | | | | | | * | * | * | * |

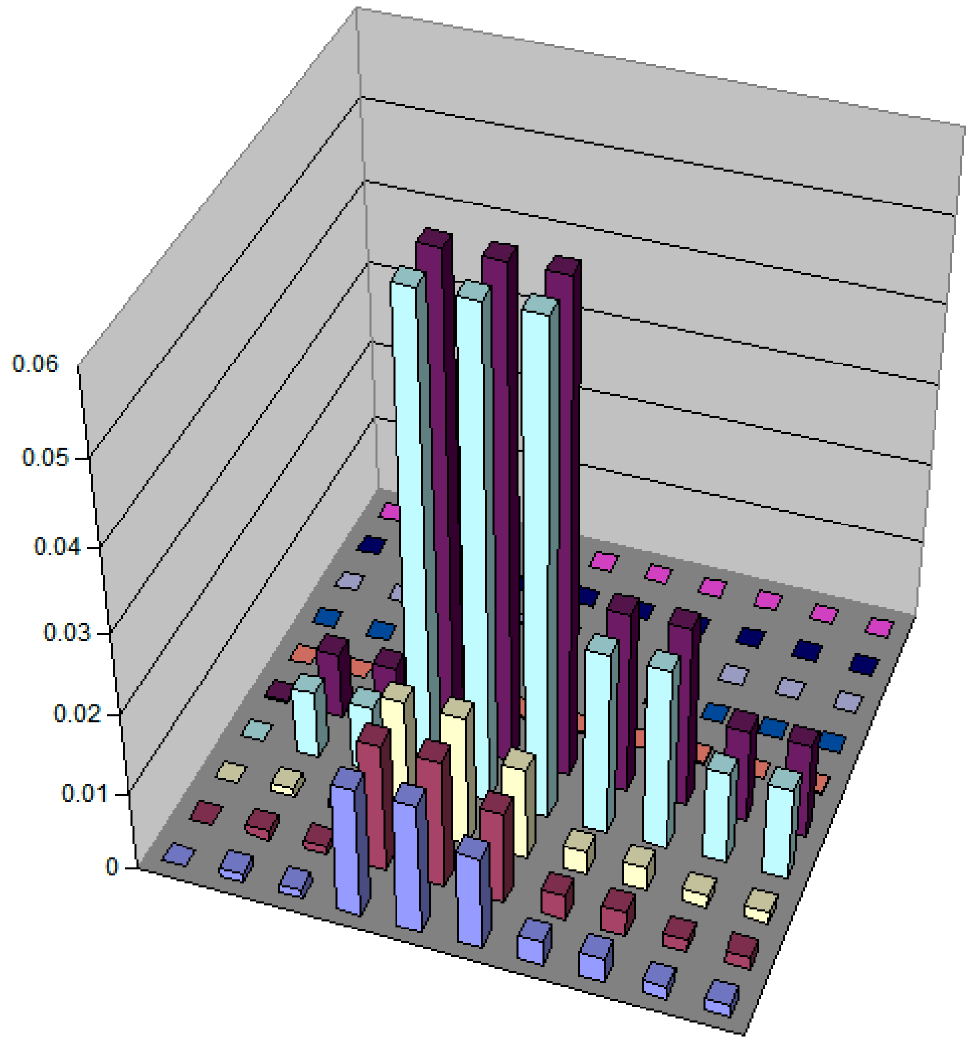
Weight: 2

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | | | | | | | | | | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |
| 8 | | | | | | * | * | * | * | * |
| 9 | | | | | | * | * | * | * | * |
| 10 | | | | | | * | * | * | * | * |

Weight: 1

## *6.4 Apply the MaxEnt formula*

- This was: $\frac{\Sigma_x \exp(-\Sigma_\mathrm{i} \mathrm{w}_\mathrm{i} \mathrm{C}_\mathrm{i}(x))}{\mathrm{Z}}$ ; slide 19 above)
- We will obtain a probability for every form.
- With other constraints not shown here added in, the graph of probability now looks like the next slide.




- Probability has been reassigned, gradiently, to a small subset of all possible forms.

## *6.5 Other Hayes/Wilson agenda items*

- Eschew a UG that has all the constraints in it; instead learn them; using a much more modest UG as starting point.
- Test the learnability implications of phonological theories (e.g. autosegmental, metrical): do they make systems learnable that would otherwise not be?

## *6.6 Sample simulation: English Onsets*

- Training set, from the CMU Online Pronouncing Dictionary:

  k 2764, r 2752, d 2526, s 2215, m 1965, p 1881, b 1544, l 1225, f 1222, h 1153, t 1146, pr 1046, w 780, n 716, v 615, g 537, dʒ 524, st 521, tr 515, kr 387, ʃ 379, gr 331, tʃ 329, br 319, sp 313, fl 290, kl 285, sk 278, j 268, fr 254, pl 238, bl 233, sl 213, dr 211, kw 201, str 183, θ 173, sw 153, gl 131, hw 111, sn 109, skr 93, z 83, sm 82, θr 73, skw 69, tw 55, spr 51, ʃr 40, spl 27, ð 19, dw 17, gw 11, θw 4, skl 1

# 6.7 Grammar fabricated: 23 constraints

| *Constraint* | *Wght* |
|---|---|
| 1. *[+son,+dors] | 5.64 |
| 2. *[+cont,+voice,–ant] | 3.28 |
| 3. *[^–voice, +ant, +strid][–approx] | 5.91 |
| 4. *[ ][+cont] | 5.17 |
| 5. *[ ][+voice] | 5.37 |
| 6. *[+son][ ] | 6.66 |
| 7. *[–strid][+cons] | 4.40 |
| 8. *[ ][+strid] | 1.31 |

| *Constraint* | *Wght* |
|---|---|
| 9. *[+lab][^+approx, +cor] | 4.96 |
| 10. *[–ant][^+approx, –ant] | 4.84 |
| 11. *[+cont,+voice][ ] | 4.84 |
| 12. *[–cont,–ant][ ] | 3.17 |
| 13. *[ ][–back] | 5.04 |
| 14. *[+ant,+strid][–ant] | 2.80 |
| 15. *[+spread][^+back] | 4.82 |
| 16. *[+cont,+voice,+cor] | 2.69 |

| *Constraint* | *Wght* |
|---|---|
| 17. *[+voice][^+approx, +cor] | 2.97 |
| 18. *[+cont, –strid][^+approx, –ant] | 2.06 |
| 19. *[ ][^–cont, –voice, +lab][+cons] | 3.05 |
| 20. *[ ][+cor][^+approx, –ant] | 2.06 |
| 21. *[+cont,–strid] | 1.84 |
| 22. *[+strid][–ant] | 2.10 |
| 23. *[–cont, –voice, +cor][^+approx, –ant] | 1.70 |

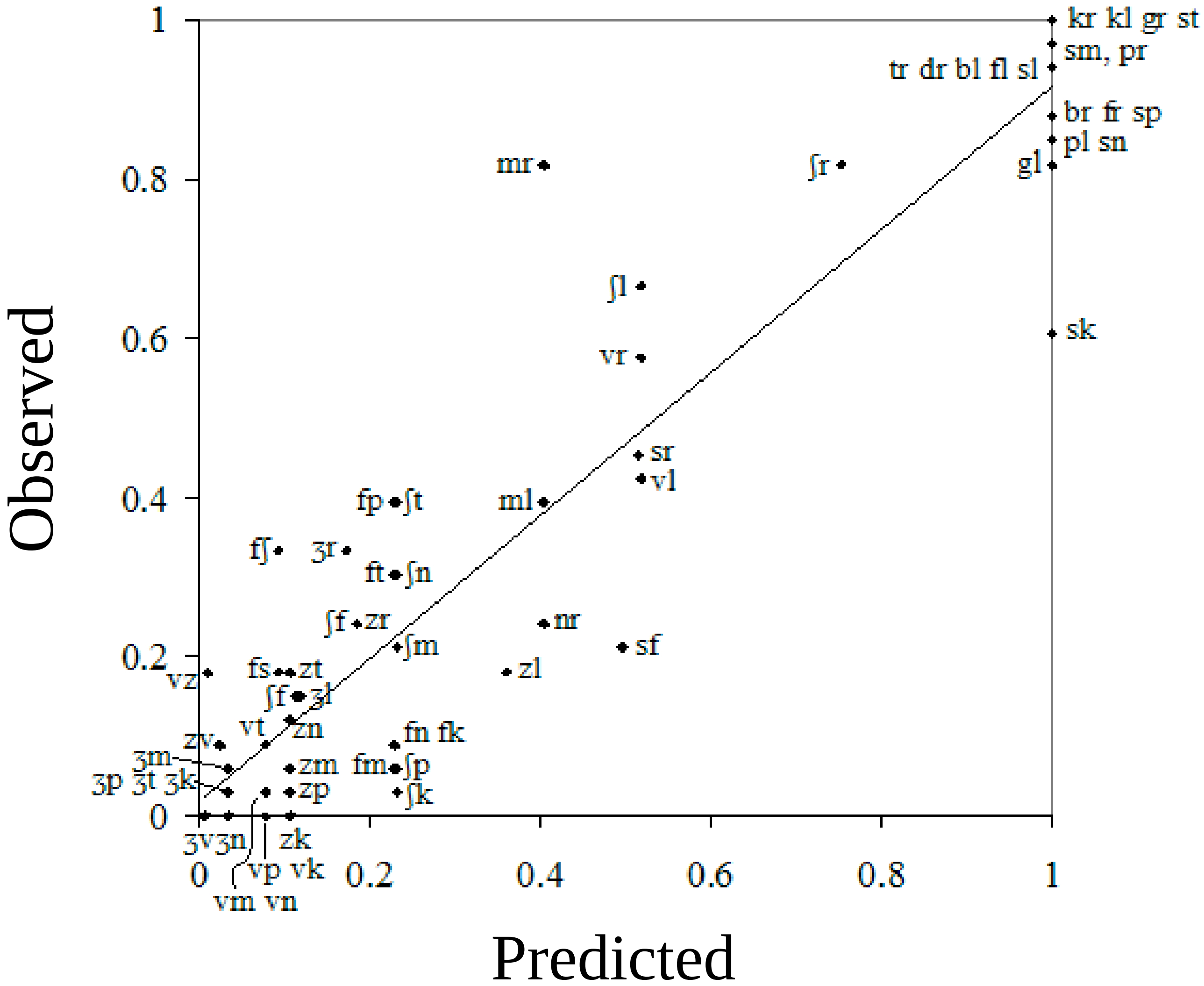
## *6.8 Testing the grammar*

- Experimental data from Scholes (1966)
- 33 subjects rated 66 monosyllabic nonce words, with ordinary syllable rhymes; independent variable was the onset.

## 6.9 Results

- Predictions of our grammar correlate well with the Scholes data, r = 0.946
- This outperforms all other approaches we tried for comparison (e.g. Coleman and Pierrehumbert 1997, *n*-gram model from Mohri 2002, Allauzen et al. 2005)

# *6.10 Scattergram: rescaled model predictions vs. Scholes data*



## *6.11 Larger scale work*

- We have analyzed the complete phonotactics of Wargamay (Australian, Dixon 1981), showing that we can fully cover at least the simpler phonotactic systems of languages.

# *6.12 Some phonotactic learning algorithms using Optimality Theory*

- Hayes 2004, Prince and Tesar 2004, Jarosz 2006
- This work assumes the standard OT approach of the **Rich Base**: a legal form is one that can derived from any input.
- Why must things be done this way? [or, why do I think this…] Because OT is based inherently on a comparison of alternatives.

# *6.13 Why the OT/Rich Base scheme may be inappropriate to phonotactic learning*

- Problems of **search space size**: we aren't just rating the forms, but any form as derived from any underlying representation.

- Previous work seem to suffer from this:

  - Hayes (2004), Prince and Tesar (2004): idealize to non-gradient learning
  - Jarosz (2006): makes a gradient system the goal, and uses the same basic strategy (maximum likelihood estimation) as Hayes/Wilson. But search space is the full set of rankings (factorial in size).

- All three: examples are schematic, not real-language.

## *6.14 Rating forms in isolation goes out on a limb*

- The idea of assigning probability just to forms (as opposed to the outputs for an input) raises many further questions—e.g., how to relate phonotactics to alternations.

# 7. Conclusions

## *7.1 Analysis of gradience in phonology cannot be taken as peripheral*

- Gradient phenomena are pervasive.
- The question of how to analyze gradience quickly moves us into the question of choice of framework, with major implications for nongradient phonology.

# *7.2 Some possible reasons for favoring a Maximum Entropy approach*

- Accuracy and trustability of its affiliated weighting algorithm.
- Perhaps: ability to assign modest probabilities to harmonically bounded candidates
- Ability to form phonotactic grammars without the use of the Rich Base principle and its accompanying search-space problem

Thank you

- Comments and afterthoughts to: bhayes@humnet.ucla.edu

# References

Allauzen, Cyril, Mehryar Mohri, and Brian Roark. 2005. The design principles and algorithms of a weighted grammar library. *International Journal of Foundations of Computer Science* 16:401– 421.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods. Journal of Memory and Language 44:568–591.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 21:43–58.

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

Coetzee, Andries. 2006. Variation as accessing non-optimal candidates. *Phonology* 26:337-385.

Coleman, John, and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonology, 49–56. Somerset, NJ: Association for Computational Linguistics.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2004) TiMBL Manual 4.0.

Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 19:180–191.

Dixon, Robert M. W. 1981. Wargamay. In Handbook of Australian languages, volume II, ed. Robert M. W. Dixon and Barry J. Blake, 1–144. Amsterdam: John Benjamins.

Eisner, Jason (2000). Review of Optimality Theory by René Kager. Computational Linguistics 26(2):286-290.

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.

Golston, Chris (1998) Constraint-based metrics. (1998) Natural Language and Linguistic Theory 16, 719-770.

Halle, John and Fred Lerdahl (1993) “A generative textsetting model,” Current Musicology 55:3-23.

Halle, John (1999) A Grammar of Improvised Textsetting. Ph.D. dissertation, Columbia University.

Halle, Morris and S. Jay Keyser (1966) “Chaucer and the theory of prosody,” College English 28: 187-219.

Halle, Morris and S. Jay Keyser (1971). English Stress: Its Form, Its Growth, and its Role in Verse. New York: Harper and Row.

Hayes, Bruce. 2004. Phonological acquisition in Optimality Theory: the early stages. In Fixing Priorities: Cstraints in Phonological Acquisition, ed. René Kager, Joe Pater, and Wim Zonneveld, 158-201. Cambridge University Press.

Hayes, Bruce. 2005. The Textsetting Problem: An Approach with Stochastic Optimality Theory. Handout for talk given at Stanford University.

Hayes, Bruce. In press. Textsetting as constraint conflict. to appear in Aroui, Jean-Louis and Andy Arleo, eds. (forthcoming) Towards a Typology of Poetic Forms. Amsterdam, Elsevier.

Hayes, Bruce and Abigail Kaun. 1996. The role of phonological phrasing in sung and chanted verse. The Linguistic Review 13, 243-303.

Hayes, Bruce and Zsuzsa Cziráky Londe. 2006 Stochastic phonological knowledge: the case of Hungarian vowel harmony. Phonology 23: 59-104.

Hayes, Bruce and Colin Wilson. 2007. A maximum entropy model of phonotactics and phonotactic learning. Ms., UCLA. Provisionally accepted at Linguistic Inquiry.

Jäger, Gerhard. 2004. Maximum entropy models and stochastic Optimality Theory. Ms., University of Potsdam.

Jäger, Gerhard, and Anette Rosenbach. 2006. The winner takes it all – almost. Linguistics 44:937–971.

Keller, Frank. 2000. Gradience in grammar: experimental and computational aspects of degrees of grammaticality. Doctoral dissertation, University of Edinburgh.
Keller, Frank. 2006. Linear Optimality Theory as a model of gradience in grammar. In Gradience in grammar: generative perspectives, ed. Gisbert Fanselow, Caroline Fery, Ralph Vogel, and Matthias Schlesewsky, 270-287. Oxford University Press.
Jarosz, Gaja. 2006. Rich lexicons and restrictive grammars – maximum likelihood learning in Optimality Theory. Doctoral dissertation, Johns Hopkins University, Baltimore, Md.
Johnson, Mark. 2002. Optimality-theoretic Lexical Functional Grammar. In Paula Merlo and Susan Stevenson, editors, The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues, pages 59–74. John Benjamins, Amsterdam, The Netherlands.
Kiparsky, Paul (1975). Stress, syntax, and meter. Language 51: 576-616.
Kiparsky, Paul (1977). The rhythmic structure of English verse, Linguistic Inquiry 8: 189-248.
Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky. 1990. Harmonic grammar: A formal multi-level connectionist theory of linguistic well-formedness: an application. In COGSCI 1990, 884–891.
Legendre, Géraldine, Antonella Sorace, and Paul Smolensky. 2006. The Optimality Theory - Harmonic Grammar connection. In Smolensky and Legendre 2006, 339–402.

Lerdahl, Fred and Ray Jackendoff (1983) A Generative Theory of Tonal Music. Cambridge, MA:  MIT Press.
Lin, Ying (2005) Learning Stochastic OT: a Bayesian approach using Data Augmentation and Gibbs sampling (pdf) (slides), ACL 05.
Maslova, Elena (to appear) Stochastic OT as a model of constraint interaction. To appear in Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, & Annie Zaenen (Eds.), Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan. CSLI publications.
McClelland, James L. and Brent C. Van der Wyk.  2006.  Graded constraints on English word forms.  Ms., Carnegie Mellon University.
Mohri, Mehryar. 2002. Semiring frameworks and algorithms for shortest-distance problems. Journal of Automata, Languages and Combinatorics 7:321–350.
Pater, Christopher Potts, and Rajesh Bhatt. 2007.  Linguistic Optimization.  Ms., University of Massachusetts.
Pater, Joe. In press. Gradual learning and convergence. To appear in Linguistic Inquiry.
Prince, Alan, and Paul Smolensky. 1993/2004. Optimality Theory: constraint interaction in generative grammar. Cambridge, Mass.: Blackwell. [Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, N.J., April 1993].

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In Fixing Priorities: Cstraints in Phonological Acquisition, ed. René Kager, Joe Pater, and Wim Zonneveld, 245– 291. Cambridge: Cambridge University Press.

Riggle, Jason. 2004. Generation, Recognition, and Learning in Finite State Optimality Theory. Doctoral Dissertation, University of California, Los Angeles.

Rumelhart, David E. and James L. McClelland, eds. (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press/Bradford Books.

Scholes, Robert. 1966. Phonotactic Grammaticality. The Hague: Mouton.

Skousen, Royal. 2002. Analogical Modeling of Language. Benjamins.

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.

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

Smolensky, Paul. 1986. Information processing in dynamical systems: foundations of Harmony Theory. In Rumelhart and McClelland 1986.

Wilson, Colin. 2007. The Luce choice ranker. Ms., Department of Linguistics, UCLA.