Class 6, 7/11/13: Bias
1. Exercise for next time
   - At end of these slides

2. Reading for next time
   ➢ On course web site
3. Where we are

- We have basic tools for doing full-scale stochastic phonology:
  - Frameworks (Stochastic OT, maxent, more)
  - Some ability to test analyses statistically with mini-maxent = logistic regression in R (more if time...)

4. For today

- Bias
- What forms of bias might be out there in the real world
  - Conservativity
  - Substantive
- How to implement bias formally in maxent
5. Looking ahead

- One more framework: noisy harmonic grammar
- Comparing the frameworks empirically: Zuraw/Hayes work on intersecting constraint families
- Other stuff I could try to squeeze in if you ask
INTRODUCTION TO BIAS
6. What is bias (in this context)?

- Any principle that keeps the learning system from achieving its objective function
- = something that wants the grammar to be some particular way, even at the cost of pure empirical accuracy
7. Why would we want to impose bias?

- More than one reason:
  - **Proper skepticism when data are few** — don’t leave to conclusions just because that would be the best-fit model.
  - **Substantive biases** — the child expects the ambient language to have certain properties, and resists learning pure data-fit grammars when these properties are contradicted.
8. Skepticism under few data

- When you hear, say, 5 examples of output A and 0 examples of output B, you can, in principle, set the weight of PREFERA extremely high, and get a superb fit.
- Let’s try this in two software programs, differing in their biases
  - Maxent in OTSoft goes immediately for a weight of 50, and achieves a huge number of decimal places of accuracy. Hooray!
  - Maxent Grammar Tool, in its default settings, computes weight of 10.1, and accuracy is to five decimal places: 0.99996.
  - Many people would prefer the more conservative value, or even more conservative.
9. **Contrasting with abundant data**

- Suppose instead that there are **one million** examples of A, 0 of B.
  - Maxent in OTSoft: same as before.
  - Maxent Grammar Tool now computes weight of 21.6; accuracy is to ten decimal places.
- So it seems sensible to want weights to be small, pending enough evidence to raise them high.
10. A common way to express suitable skepticism in maxent: the Gaussian Prior

- "Prior" = prior expectation, updated by data.
  - This comes straight out of Bayesianism, a hugely important concept of our time.¹
  - The prior is your default expectation about reality, prior the input of belief-altering data.

- Gaussian: see later

- For every constraint we stipulate two values:
  - $\mu = \text{the preferred value}$
  - $\sigma = \text{the degree to which the weight is "willing" to diverge from } \mu \text{ under the pressure of learning data}$

¹ I rather enjoyed the recent bestseller *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy* by Sharon Bertsch McGrayne.
- For conservativeness against few data, we would posit $\mu = \text{zero}$.
- For a constraint that we think is a priori ranked highly in UG (Markedness? OO-Faithfulness?), we posit a high value of $\mu$. 
11. Another approach to bias

- Source: Wilson (2006), the pioneering paper in this area
- Set all the $\mu$ to zero.
- Let the constraint vary by $\sigma$.
- Then: a “strong” constraint in UG is one that easily becomes powerful under the influence of data.
12. What the Gaussian prior amounts to mathematically

- The objection function previous was:

\[ \ln(\text{probability(data under model)}) \]

- Now it is:

\[ \ln(\text{probability(data under model)}) - \sum_{j=1}^{\mu} \frac{(w_j - \mu_j)^2}{2\sigma_j^2} \]

where the second term is the Gaussian prior.

- Computing the prior value, in words:
  - Take each weight and find its distance from its \( \mu \).
  - Square the result
  - Divide by twice its \( \sigma \) squared
  - Sum across all constraints
• So, since you’re maximizing the objective function, and there’s a minus sign …
  ➢ you are minimizing distance of each weight from its $\mu$
  ➢ If $\sigma$ is small, you’re dividing by a small number, so penalty is big.
13. The Gaussian prior and avoiding infinite weights

- When I said the search space is a Magnificent Dome, I lied a bit…
- Suppose you have
  - Candidate A, frequency 1000
  - Candidate B, frequency 0
  - Constraint1 penalizes B but not A.
  - Constraint2 is indifferent
- Then: best grammar gives an infinite weight to Constraint1
- So the “dome” peaks at infinity.
- The Gaussian prior makes the best solution a (large) finite weight.
- OTSoft, with no Gaussian prior, sets a user-specified maximum (default 50).
14. How to specify a prior in the Maxent Grammar Tool

- Make a little file with a name like Prior.txt.
- Each line: Constraint name, tab, $\mu$, tab, $\sigma$.
- Minor bug: you have to put a minus sign in front of your mu’s if you want them to be positive; sorry.
- When you run the program, access Prior.txt with the middle button, labeled “Open constraints.”
15. Back to our toy example, with a prior

- We’ll do a million learning cases again.
- But we’ll set sigma at 1, instead of the program default of 100000.
- Accuracy descends to just four decimal places — still high, since a million examples is a great deal of data.
16. Coming up

- Applications of the prior
  - Evidence that human learners are skeptical in the face of insufficient data.
  - Evidence that human learners actively prefer to give stronger weights to certain constraints.
CONSERVATIVITY BIAS
17. Source


➤ = reading for this time
18. Theme

- Martin studies cases where allowing weights to go very high would achieve a perfect fit.
- Modeling with a Gaussian prior holds back the weights.
- This leads to detectible empirical language-acquisition effects.
19. The basic distribution of geminate consonants in English

- No geminates (long/double consonants) within a morpheme: there can be no minimal pair [hæpi]/[hæppi].

---

2 I find it amusing that in the first moment of January 1 one is allowed to shout: [ˈhæppi ˈnu: jʊ], but that’s part of the expressive phonological system.
20. The statistical distribution of geminates

- Martin discovered that geminates are less common than would be expected by chance—that is, there are not as many words like *bookcase* as expected:
21. Martin’s method

- Martin is a frequent user of the **Monte Carlo** method in statistics.
- This is a delightfully simple method that can handle all sorts of tricky cases.
- Just
  - Randomly simulate what you want.
  - Do it over and over again to create a distribution.
  - See where reality falls within your distribution.
22. Monte Carlo for English compounds

- Pick Word1 at random.
- Pick Word2 at random.
- Concatenate
- See if the resulting compound has a geminate.
  - E.g. if you picked *book* as Word1 and *case* as Word2
- Do this 4,578 times.
- Count the geminates.
- Do the above thousands of times to create a distribution of this value:
  - Number of geminates/4,578

23. Reality is far below simulated reality
Moreover (bottom graph): phonologically legal clusters are overrepresented.
24. Upshot

- Geminates are legal in compounds but people don’t seem to like them!
- Why so?
- Martin:
  - the grammar characterizes them as less phonotactically probable
  - lower probability is a factor in the adoption and spreading of novel words (J. Pierrehumbert talk/course, this Institute)
25. Explaining the underrepresentation effect

- We will redo Martin but in our own naïve terms with the class software.
- Key idea:
  - Conservativity bias keeps the learner from attributing an effect to a single (in fact, to us, true) case, but must distribute the effect around the grammar, creating minor errors.
26. Constraints

- **ANY GEMINATE**
  - Note: plenty of languages *actively degeminate* when morpheme concatenation creates a geminate. Google Scholar search: “degemination - degermination”

- **STEM-INTERNAL GEMINATE**
  - True for English; stem-bounding is common for phonological constraints.
• *NULLPARSE
  ➢ assumes that “no word emerges from the word formation component” is a “candidate”
  ➢ it is penalized; we’d *like* to make new words
  ➢ If it wins, the speaker shuts up or chooses to say some other word.
## 27. Candidates and violations

<table>
<thead>
<tr>
<th></th>
<th>*ANY GEMINATE</th>
<th>*STEMINTERNALGEMINATE</th>
<th>*NULLPARSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>pp</td>
<td>pp</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NullParse</td>
<td>100 0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>p+p</td>
<td>p+p</td>
<td>100 0</td>
<td></td>
</tr>
<tr>
<td>NullParse</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
28. “Rational” ranking/weighting

- This matches data as well as possible
- *Stem-internal geminate is super-strong, dominating
  *NullParse
- *NullParse is strong enough to dominate *Any Geminate
29. This weighting is found by Maxent Grammar Tool with its default settings

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Mu</th>
<th>Sigma</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>*ANY GEMINATE</td>
<td>0</td>
<td>100000</td>
<td>0</td>
</tr>
<tr>
<td>*STEMINTERNALGEMINATE</td>
<td>0</td>
<td>100000</td>
<td>28.4</td>
</tr>
<tr>
<td>*NULLPARSE</td>
<td>0</td>
<td>100000</td>
<td>14.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Candidate</th>
<th>Observed Freq.</th>
<th>Observed Proportion</th>
<th>Predicted proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>pp</td>
<td>pp</td>
<td>0</td>
<td>0</td>
<td>0.0000000 6</td>
</tr>
<tr>
<td></td>
<td>NullParse</td>
<td>1000</td>
<td>1</td>
<td>0.9999999 4</td>
</tr>
<tr>
<td>p+p</td>
<td>p+p</td>
<td>1000</td>
<td>1</td>
<td>0.9999999 2</td>
</tr>
<tr>
<td></td>
<td>NullParse</td>
<td>0</td>
<td>0</td>
<td>0.0000000 8</td>
</tr>
</tbody>
</table>

30.
31. Let’s try installing a strict Gaussian prior — penalize large weights

- We’ll make the prior strict by setting $\mu$ at zero and $\sigma$ at 0.1.
- Putting this in the `Prior.txt` file and running it, we get:

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Mu</th>
<th>Sigma</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ANY GEMINATE</em></td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td><em>STEMINTERNALGEMINATE</em></td>
<td>0</td>
<td>0.1</td>
<td>4.3</td>
</tr>
<tr>
<td><em>NULLPARSE</em></td>
<td>0</td>
<td>0.1</td>
<td>1.9</td>
</tr>
<tr>
<td>Input</td>
<td>Candidate</td>
<td>Observed Freq.</td>
<td>Observed Proportion</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>pp</td>
<td>pp</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>NullParse</td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td>p+p</td>
<td>p+p</td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>NullParse</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
32. Can we make sense of this?

- The geminates in compounds are losing to the null parse 13% of the time
  - This gives Junior pause when he seeks to utter a new compound, say (in the bath) “Daddy, I want my tub bottle.”
- What about the cases of legal stem-internal geminates?
  - Inference: for a period of time Junior would not reject stem-internal geminates.
  - Perhaps true, if he is misinterpreting some bookkeeper compounds as such.
  - Once Junior has learned more data and overcome the prior, the language has already gotten a lexicon where bookkeeper compounds are underrepresented.
• See Martin for other views.
33. Many more data overwhelm the prior

- Let’s try it once more with the same prior but a million data of each type.

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Mu</th>
<th>Sigma</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>*ANY GEMINATE</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>*STEMINTERNALGEMINATE</td>
<td>0</td>
<td>0.1</td>
<td>15.7</td>
</tr>
<tr>
<td>*NULLPARSE</td>
<td>0</td>
<td>0.1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Candidate</th>
<th>Observed Freq.</th>
<th>Observed Proportion</th>
<th>Predicted proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>pp</td>
<td>pp</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>NullParse</td>
<td>1000000</td>
<td>1</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>p+p</td>
<td>p+p</td>
<td>1000000</td>
<td>1</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>
• Indeed, in any sort of Maxent learning “the data eventually overwhelm the prior”
• We must suppose that in real life:
  ➢ the prior is matched against a corpus roughly equivalent to childhood experience
  ➢ learning ceases, perhaps through loss of plasticity
34. Upshot

- Underrepresentation of geminate-compounds is expected as a “spillover” of illegal geminates in stems, IF
  - we implement Conservativity-biase with a weight-penalizing Gaussian prior.
  - learning data are not quite enough to settle the weights onto the empirically appropriate constraints.
35. Similar findings by Martin in Navajo compounds

- Within a word, sibilants *must* agree—affixes even alternate.
- In compounds, they tend to agree (if in adjacent syllables)
36. Similar finding in Turkish compounds

- Vowels within a stem tend strongly to agree in backness
- Vowels within a lexicalized compound tend—less strongly but still significantly—to agree
37. Generalizing further

- Find cases where there is
  - A specific constraint — accurate for the language; accounts for gap
  - A more general version — not accurate; penalizes the legal
- Prediction: there should be an empirical “negative halo”; underrepresentation of forms covered by the more general version.
- Reason: Conservativity bias forces the general constraint to take over some of the work of the specific one.
SUBSTANTIVE BIAS
38. Substantive bias

- This is a major topic in current phonological theorizing — work of Joe Pater, Karen Jesney, Anne-Michelle Tessier, Elliott Moreton, Colin Wilson, Adam Albright, many others.

- The hypothesis:
  - Phonology is learned not simply by induction with constraints over data (perhaps conservative induction) but with prior expectations by the child about what a phonology should be like.
  - Theory is engaging with experimental data, notably artificial-grammar learning experiments.
39. Substantive bias can be taxonomized following the natural architecture of OT

- **Faithfulness**: biases discourage the learning of phonological alternation, particularly phonetically salient alternation.

- **Markedness**: biases discourage the learning of patterns that create marked surface forms.
  - Research so far suggests, to me, that Faithfulness-based naturalness patterns are extensive; Markedness-based ones dubious and hard to find.

- Let’s try Faithfulness first.
40. My favorite form of substantive bias

- A preference for similarity among the allomorphs of a single morpheme: minimize alternation.
- There is some evidence (e.g. Zuraw (2007, Language)) that this preference is based on raw phonetic similarity; not abstract feature-based similarity.
- If this is so, then the Faithfulness we should consider is Output-Output Faithfulness, not Input-Output Faithfulness (i.e. to an abstract underlying representation).
41. Theoretical apparatus

- Steriade’s P-map theory

- Zuraw’s P-map-based *MAP constraint family*
42. *MAP constraints (Zuraw 2007)

- *MAP(x, y) assesses a violation to a candidate if a segment belonging to natural class $x$ in the input is mapped to a corresponding segment in natural class $y$ in the output.
- This kind of constraint is more flexible than classical OT correspondence, since there is no requirement that $x$ and $y$ differ by just one feature.
- It becomes important when we consider phonologically unusual alternations.
43. Default rankings of *MAP

- Appealing to phonetic substance, *MAP constraints are assigned a default ranking as follows:
  - *MAP constraints banning changes that cover a larger perceptual distance are assigned a default ranking higher than constraints banning smaller changes.

- This ranking preference is taken to be a learning bias in UG.
  - Given sufficient evidence in the ambient language, it is possible for learners to subvert the default rankings, but this is harder.
44. Rankings in *Map are based on the P-map

- Zuraw draws on earlier work by Steriade (2001, 2008) on the **P-map**.
  - = a compilation of the language learner’s phonetic experience concerning the perceptual distance between forms.
  - From a Steriade ms. on line; font size = perceptual distance.
<table>
<thead>
<tr>
<th>Obstruent voicing</th>
<th>V_</th>
<th>C_</th>
<th>V_</th>
<th>V_</th>
<th>V_</th>
<th>C_</th>
</tr>
</thead>
<tbody>
<tr>
<td>p/ b</td>
<td>p/</td>
<td>p/</td>
<td>p/b</td>
<td>p/b</td>
<td>p/b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>t/ d</td>
<td>t/d</td>
<td>t/d</td>
<td>t/d</td>
<td>t/d</td>
<td>t/d</td>
<td>t/d</td>
</tr>
<tr>
<td>k/ g</td>
<td>k</td>
<td>k</td>
<td>k/g</td>
<td>k/g</td>
<td>k/g</td>
<td>k/g</td>
</tr>
<tr>
<td>g</td>
<td>g</td>
<td>g</td>
<td>g</td>
<td>g</td>
<td>g</td>
<td>g</td>
</tr>
</tbody>
</table>
- It is assumed that language learners consult the P-map in assessing preferred rankings for *MAP constraints.
46. The analytical challenge ahead

- Use experimental data to make a real, quantitatively-explicit P-map.
- Use this P-map to impose plausible biases in phonological learning.
47. Backdrop to *MAP-cum-P-map theory

- Essential prediction: alternation prefers to be phonetically minimal; proposed long ago (Vennemann 1972, Kiparsky 1978/1982:65)
48. There’s a phonological literature backing up a bias for phonetically-minimal alternation

  - a 15-year research program on “analogical” change
  - languages can even create new phonemes when this reduces the distance of phonological alternation

- Language acquisition
  - I’m not sure how common this is, but here is one relevant case:

Children are able to innovate sequences that are illegal in the target language, in the interest of maintaining output-to-output correspondence. This was observed by Kazazis (1969) in the speech of Marina, a four-year-old learning Modern Greek. Marina innovated the
sequence *[xe] (velar consonant before front vowel), which is illegal in the target language. She did this in the course of regularizing the verbal paradigm: thus ['exete] ‘you-pl. have’ (adult ['eçete]), on the model of ['exo] ‘I have’.3

- Fleischhacker (2001), Shademan (2002) [both on UCLA dissertations page]
  ➢ epenthesi s favors location that changes the stem least saliently
- Zuraw (2007, cited above)
  ➢ infixation favors location that changes the stem least saliently

---

• Lofstedt (2010, UCLA diss., at UCLA website)
  ➢ allomorph (irregular) phonological variation in Swedish; leveling happens where

• Ancient history
  ➢ the widely-accepted-but-vague “phonetic similarity” requirement in phonemicization ([h]-[ŋ], [pʰ]-[t]/[tʰ]-[k]/[kʰ]-[p], e.g. Hayes 2009 textbook)
49. The source paper on bias as implemented in maxent grammars


- This was
  - a pioneering effort to study naturalness through artificial-language learning (palatalization of velars preferred before high vowels)
  - an early use of artificial-language experiments to explore learning bias
  - the first formal implementation of Steriade’s P-map
  - first use of maxent grammars with Gaussian priors to model biased acquisition:
• Note: Moreton and others have questioned whether Wilson’s experiments fully demonstrate what he is arguing for.

50. Next time’s example

• I’ll do recent work by a recent UCLA Ph.D. that I think shows more unequivocal results using Wilson’s methods.

51. Exercise

- Use the Maxent Grammar Tool and the J. White input files on the course web site to check the following question: what is the trade-off between changing the training data, and changing sigma? For example, is doubling the amount of training data equivalent to cutting sigma by half?