Class 4, 4/9/2020: More on frameworks; bias

1. Assignments

- Read:
  - Colin Wilson and Gillian Gallagher (2016) Accidental Gaps And Surface-based Phonotactic Learning: A Case Study Of South Bolivian Quechua. *Linguistic Inquiry* 49:610-623.
  - On course web site
- Continue homework on medial clusters.
  - Due in a week, Thursday April 16.
  - Feel free to consult me, remembering that my office hours are by appointment (ask during break or by email).

2. Outline for this class

- A bit on Stochastic OT
- Noisy Harmonic Grammar as an alternative to MaxEnt
- Running NHG on a spreadsheet
- UG bias and how to implement it in grammar learning

3. Coming up

- The big debates over MaxEnt and NHG: harmonically bounded winners, factorial typology size
- If time: the Observed/Expected fallacy and how to avoid it in MaxEnt
- Acquisition

THE FRAMEWORK BAZAAR

4. Sources

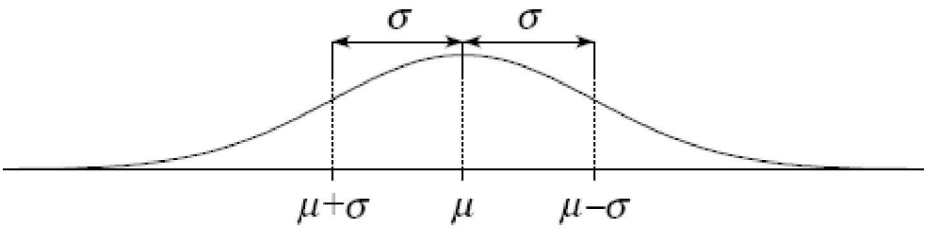
- Boersma dissertation (1998)
- I sought to collaborate with Boersma when I heard about this, resulting in a joint paper applying it to phonology (Boersma and Hayes 2001, *LI*)

5. Quickie version

- Every constraint has a number that represents its strength.

(high ranked) (low ranked)

- Jiggle each such number whenever you use the grammar, assigning a bit of Gaussian noise.



- Once you have jiggled, sort the resulting values to get a complete constraint ranking, follow the standard OT selection procedure to get a winner.

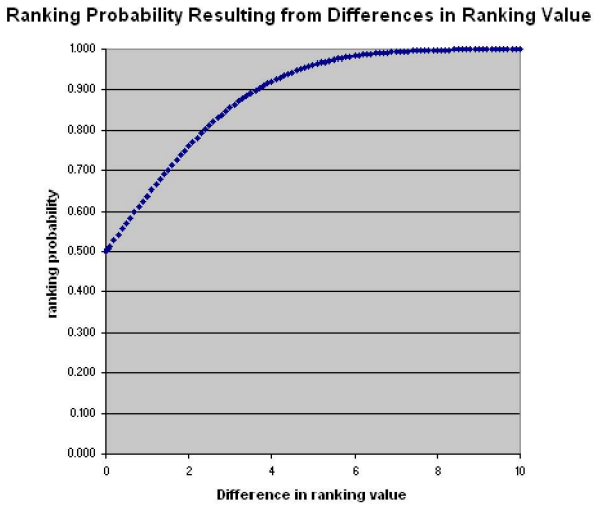
6. The function relating ranking-value-difference to probability-of-the-more-likely winner

- You can do this in Excel

http://www.linguistics.ucla.edu/people/hayes/GLA/RankingValuesToProbabilities.xls

- The spreadsheet uses the Excel NORMDIST( ) function.

Difference in ranking value	*Probability higher outranks lower*
0	0.5
0.1	0.51
0.5	0.57
1	0.64
5	0.96
10	0.9998
50	1.00000000



- [☞ comment on the shape of this curve]

7. Evaluation

- See Hayes reading, and work cited there, for the poor ganging ability of this theory, and the diagnosis.
- Finding an algorithm to set the weights (a.k.a. ranking values) has proven to be very problematic.
  - No provably convergent algorithm exists.
  - The leading one, the Gradual Learning Algorithm, behaves erratically; fails on simple grammars (Pater (2008, *LI*), Magri and Storme *LI* 2019), and often sends the weights off toward infinity as no convergence occurs. Plenty of frustration in my own personal history as a user.
- The theory was defended and elaborated, persistently and brilliantly, for years by Giorgio Magri, — who, whoever, seems to have moved on to Noisy Harmonic Grammar as his favorite.

8. Noisy harmonic grammar

- References:
  - Boersma, Paul, and Joe Pater. 2008/2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. Amsterdam and Amherst, MA: University of Amsterdam and University of Massachusetts ms. Rutgers Optimality Archive. Published 2016 in John McCarthy and Joe Pater, *Harmonic Grammar and Harmonic Serialism*
  - Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker (2010). Harmonic Grammar with linear programming: From linear systems to linguistic typology. *Phonology* 27, 77-117. (non-stochastic version)
- This is a lot like MaxEnt; again you calculate a Harmony score for every candidate.
- But you jiggle the harmony scores stochastically, deriving a winner for each evaluation time, just like in Stochastic OT.

9. Many varieties exist

- See
  - Bruce Hayes (2017) Varieties of Noisy Harmonic Grammar. *Proceedings of the 2016 Annual Meeting in Phonology*, USC.
- Specifically: where do you put the noise?
  A. On the constraint weights (= classical version)
  B. On the asterisks
  C. In the tableau cells (Goldrick and Daland, *Phonology* 2009)
  D. On the harmony values (behaves very much like maxent)

10. Assessment

- I personally feel this framework is in contention:
  - Performs about as well in practice (I suspect) as maxent.
  - No proof of convergence for Boersma and Pater's learning algorithm, but I have never seen it misbehave as the Stochastic OT algorithms often do.

- Is intuitive just like MaxEnt (as it is also a form of stochastic Harmonic Grammar).

11. Calculating probabilities of outputs in NHG: technique

- Crude: run the random number generator thousands of times, and count what you got.
  - This will not return exact values — often useful in diagnosis
  - The error is often fairly substantial unless you sample forever
- Analytic solutions (as in MaxEnt)
  - see below

12. Some online software that can carry out the sampling strategy

- Praat (Boersma and Weenink)
- OTHelp (UMass; https://people.umass.edu/othelp/)
- OTSoft (Hayes website; https://linguistics.ucla.edu/people/hayes/otsoft/)

13. The analytic solution — two viable candidates only

- Assume for simplicity the version of the theory with candidate-level noise = (9D)
- Cand1 and Cand2 each have a base harmony, which then gets perturbed by noise
  - Let this be a Gaussian with a particular standard deviation, say, 1.
- The base harmony for each of Cand1 and Cand2 is μ, the mean (hence peak) value of the Gaussian distribution that constraint — overlapping Gaussians.
- If Cand1 is to win, it must have lower harmony penalty than Cand2 at evaluation time.
- This is the probability that when we sample from Cand's Gaussian, we get a lower value than when we sample from Cand's Gaussian.
- Beautiful, Googleable math: the difference between two Gaussians is itself a Gaussian
  - Mean: the difference of the means
  - Standard deviation: root mean square (square root of sum of squares) of the standard deviations
- So we look at the probability that the difference distribution, Cand1 − Cand2, **is less than zero**.
- This can be done in Excel …

14. Let's do this for Claire's English stress data

- See spreadsheet
- Key point that emerges (MaxEnt as well): you don't need a conjoined constraint (stats: interaction term).
  - Three parameters (weights), four observations (winning frequencies of four candidates), so something (a little!) is at stake.

15. Generalizing the method to more than two candidates

- A candidate wins if it beats all others.

- So we would have to compute the product of the probability of all these wins, to get the probability of an outright first-place win.

16. Generalizing the method to more complicated forms of NHG

- We just have to build up the Gaussians step by step (sum of Gaussians works just like difference of Gaussians), to the point that we know the distribution of Harmony for each candidate.
- Or we could collapse in a heap and just do sampling …

BIAS

17. Hard UG has been a tough row to hoe

- I had no idea when I was told about Linguistic Universals as a teenager how many proposed "hard" universals would bite the dust!
- Examples
  - Wh-Island Constraint, Complex NP Constraint (Goodluck and Rochemont 1992:6-9)
  - Coordinate Structure Constraint (e.g. Oda 2017)
  - Unbounded stress patterns always default-to-opposite (Hayes 1995:33)
  - No "even iambs" (Altshuler 2009)
  - No myopia (Copperbelt Bemba, Kula and Bickmore 2015)
  - No majority rules Warlpiri vowels take a vote on whether to be all front or all back, when harmony-irregular words get regularized (Zymet and Bowler on Warlpiri.).
  - See Hayes and Jo, on Hayes web site, to get the citations here

18. Evidence is accumulating for a soft UG — biases

- Main source:
  - **non-veridical learning**
  - either in the experimental participant's life experience (readings), or in an artificial grammar learning experiment

19. How to obtain testing cases for nonveridical learning in real life

- Language do obtain unnatural patterns through the accidents of history.
- English fricative voicing is almost entirely in monosyllables; why?
  - The processes that created it generally created monosyllables.
  - It's not so productive, so newly-arriving polysyllables[1] tend not to undergo.
- Diachrony is the most perfect source of stupid explanations for things (though you have to be smart to do diachrony well :=))

[1] Except BH's *epitaph*

- The fricative-voicing is just the tip of the iceberg: a whole research tradition on "crazy rules" and "unnatural phonology"
  - See Hayes and White (*Phonology* 2017) for a quick overview.

20. A subset of the literature on non-veridical learning

- Probably the ur-reference:
  - Wilson, Colin (2006) Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30 (5), 945-982
  - His UG principle is the **P-map** (Steriade, Zuraw, more later on): avoid alternation when it is phonetically salient.
  - ki ~ tʃi is less phonetically salient than ke ~ tʃe
  - Artificial grammar experiment: train on ke ~ tʃe, generalizes to ki ~ tʃi, but not the other way around.
  - Various cans of worms; see for instance the critique by Elliott Moreton (2008) "Analytic bias and phonological typology"; *Phonology* 2008
- Readings: Becker, Nevins, and Levine on the special faithfulness devoted to initial syllables.
  - Same authors have done this for other languages.
- (2009) Bruce Hayes, Kie Zuraw, Peter Siptar, and Zsuzsa Londe. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85: 822-863.
  - Vowel harmony is triggered primarily by vowels (yay!)
  - There are also weird, arbitrary, but statistically significant consonant effects. ("use front suffixes when the stem ends in a sibilant").
  - These are treated less seriously in a wug test than the more natural vowel effects.
- Work of James White on **saltation**
  - White, J. (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. Language, 93(1), 1–36.
  - Hayes, B. & White, J. (2015). Saltation and the P-map. Phonology, 32(2), 1–36.
  - White, J. & Sundara, M. (2014). Biased generalization of newly learned phonological alternations by 12-month-old infants. Cognition, 133(1), 85–90.
  - White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. Cognition, 130(1), 96–115.
  - Basic scheme: p → β, b → b intervocalically. This is unbelievable to undergraduates and also to babies; they want b → β as well. Same explanation as in the Wilson study.

21. Toward modeling such effects: the Gaussian prior

- The idea for this comes from Wilson (2006).
- The technical concept, which is standard computer science, is presented clearly in Goldwater and Johnson's (2003) paper (which reintroduced MaxEnt into phonology).

- This is the formula for the objective function, maximized in finding the best weights.

$$\log \mathrm{PL}_{\bar{w}}(\bar{y}|\bar{x}) - \sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2}$$

this part is the likelihood of the data; log probability under a batch of weights w of the data y given inputs x.

This part is the Gaussian prior

22. Calculating the prior

- It is a penalty, subtracted from the likelihood.
- It will cause the weights to differ, somewhat, from those that maximize the likelihood.
- Each mu is the "favorite" value for constraint weight w_i, since if the constraint weight is at the value of mu, there will be no penalty (per formula)
- Each sigma is a value of "flexibility": how willing is the weight to deviate from its ideal value?
  - N.B. this is inverted, because it is in the denominator.
  - Sigmas like 1 are powerful; sigmas like 100000 are virtually absent.

23. Aside — why the name?

- The expression above is in the domain of log probability (the usual computational unit of MaxEnt).
- But do the math to convert it into probability (i.e. take *e* to the whole thing): you get a Gaussian curve, with □ as its mean and σ as its standard deviation.

24. There are other priors

- You don't have to square the deviation.
- Such a linear prior tends to favor grammars in which few constraints do the work.
- Gaussian prior disfavors high weights, so it tends to make constraints share the work.

25. Why is it called a "prior"?

- This comes from Bayesian probability theory, which is about how you update your beliefs based on data.
- The prior is the starting point, updated once you encounter data.
- So here, the mu's form our a priori belief about what the constraint weights are.
  - And thus mu 's are in principle a way to implement UG.

26. A convenience- virtue of a prior

- Suppose a constraint is never violated in winners — top stratum in traditional OT.
- As we've seen, the higher the weight we give it, the harsher the penalty on violating candidates.
- But remember, *maxent never reaches zero probability*.
  - This is a design feature, not a bug! Recall that ever more evidence is needed to approach certainty.
- Things go badly with most computational equipment if we let weights approach infinity.
- So a very modest prior is useful in preventing crashes — even 100,000 suffices.

27. Determining the prior in modeling work

- Choice 1: is constraint strength carried out by varying m's, or σ's, or both?
- Wilson (2006): highish mus, weak constraint have high σs and can thus be "demoted" easily.
- James White oeuvre: σ always the same; mus directly reflect constraint strength.

28. Sigma and experience

- Note that the prior stays the same no matter how many data you have.
- But with acquisition, more and more data pile up.
- You can mimic acquisition either by gradually adding (artificially?) data, or shrinking sigmas.

29. Making the mu's rigorous

- Ideally, they come from somewhere, not the investigator's head.

30. General notions of constraint strength

- Output-to-output correspondence is stronger than Markedness.
  - Because children are believed to say impossible things to make paradigms uniform.
  - From me (2004) "Phonological acquisition in Optimality Theory: the early stages. In Kager, Rene, Pater, Joe, and Zonneveld, Wim, (eds.), Fixing Priorities: Constraints in Phonological Acquisition. Cambridge University Press.

> Another source of evidence on this point comes from observations of children during the course of acquisition: 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 4-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 regularising the verbal paradigm: thus ['exete] 'you-pl. have' (adult ['eçete]), on the model of ['exo] 'I have'.

31. Phonetically based priors

- Wilson, White both used confusion matrix data to derive measures of similarity, which then map onto priors.
- Goal is to punish salient alternation.
- White uses a very easy method: find the weights for each IDENT(feature) constraint in a maxent grammar that predicts confusion rates.

32. A toy example: Modeling Moore-Cantwell's experimental data

- Scheme:
  - We have a lexicon of trisyllabic words with patterns involving heavy penult, final [i].
  - We assume this is representative of the learning data encountered by Claire's Mechanical Turk participants.
  - But their intuitions are less extreme than the lexicon.
  - Let us assume donkey-like reluctance to take the data seriously. [☞ Why else might the experimental patterns come out "weaker" than the lexical patterns?]
  - We model this with the same prior, mean zero, sigma to be determined, for all constraints.
- Is this legit in such a tiny cases?
  - We now have four data points, four parameters.
  - Yet the theory cannot compute any set of values, so there is in principle some sort of legitimacy to this.
- Try the spreadsheet

33. Work that has tried to use the bias method on a more realistic scale

- Wilson, Hayes et al., White, cited above