Class 4, 3/11/2018: Bias II; Wilson/Obdeyn on model evaluation

1. Assignments

- Read:
  - Colin Wilson and Marieke Obdeyn (2009 ms.) Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay.
- Continue homework on medial clusters.
  - Due next class, Monday April 16.

FORMALIZING BIAS WITH MAXENT: THE GAUSSIAN PRIOR

2. Review of the log-likelihood objective function

- Recall the normal objective function: maximize the likelihood of the data
  - = probability assigned to the set of winners by the grammar.
  - This is normally done by maximizing log likelihood, to avoid software crashes.
  - This seems to work beautifully, often providing perfect fits in simple problems.

3. Bias

- We think a priori that the weights are likely to be *thus*: …
- Data causes us to revise our conception of the weights, but the final outcome might still be influenced by our prior conceptions.

4. Formalizing bias in maxent

- From Goldwater and Johnson's (2003) paper, reintroducing 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 old: the log-likelihood of the data; log probability under a batch of weights *w* of the data *y* given inputs *x*. (bars evidently mean "a bunch of them; all of them")

This part is the **Gaussian prior**.

5. Calculating the prior

- It is a penalty, subtracted from the likelihood (which is a credit).

- It will cause the weights to differ, somewhat, from those that merely maximize the likelihood.
- Each μ is the "favorite" value for constraint weight w_i, since if the constraint weight is at the value of μ, there will be no penalty.
- Each σ 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.
  - σs like 1 are powerful; σs like 100000 are virtually absent.

6. Why is it called a "prior"?

- This comes from Bayesian probability theory, which is about how you should rationally update your beliefs based on data.
- The key element is Bayes' Theorem:

$$P(H) = \frac{P(D|H) \times P(H)}{P(D)}$$

"To update your estimate of the probability of the hypothesis, multiply the likelihood of the data under the hypothesis, and divide by the probability of the data."

Probability of the data = weighted sum of probability of the data under all hypotheses.

- In Bayes' Theorem, P(H) is the **prior**; i.e. what you thought about the hypothesis before you encountered the data.
- A number of papers give Bayes' Theorem as their first formula, then develop it by plugging locally relevant hypotheses and data.

7. Why is it called a "Gaussian prior"?

- Let us focus on the prior term, from the right side of Goldwater and Johnson's formula:

$$-\sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2}$$

- For simplicity let us consider the case of one constraint, so big sigma and subscripts can go:[1]

$$-\frac{(w - \mu)^2}{2\sigma^2}$$

- We are subtracting in log probability, which would be division in probability.
- Since the probability domain is the "real" one, let's get back to it by taking *e* to the value:

[1] In the general case, the terms will add in the log domain, thus multiply in the probability domain, which is correct.

$$e^{-\frac{(w-\mu)^2}{2\sigma^2}}$$

- But as the Internet tells us, this is the Gaussian distribution itself, with mean μ and standard deviation (you also have to multiply by the factor shown):

Distribution	Functional Form	Mean	Standard Deviation
Gaussian	$f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-a)^2}{2\sigma^2}}$	a	σ

8. Interpreting μ's and σs

- Now that we know that this is a Gaussian distribution, we can interpret the μ's and σ's for their meaning.
- The μ's form our a priori belief about what the constraint weights are.
  - The *most probable* a priori weight is where the maximum of the Gaussian curve occurs, at μ.
- The σ's say how *uncertain* we are about where the weights are:
  - What is the probability of other values of the weight, falling some number of standard deviations away from μ?

9. A purely-computational virtue of a prior

- Suppose a constraint is never violated in winners — top stratum in traditional OT.
- The higher the weight we give it, the harsher the penalty on violating candidates, and so the better the grammar.
- 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.
- So there is nothing to keep an algorithm from sending the weight off toward infinity.
- Things go badly with your computational equipment if this happens.
- So a very modest prior is useful in preventing crashes.
  - A very minimal one, like σ = 100,000, suffices for this purpose and will not distort the probabilities you derive other than infinitesimally.
  - When lazy, we have sometimes imposed an arbitrary upper weight limit instead.

10. Determining the prior for UG modeling

- Is constraint strength carried out by varying μ's, or σ's, or both?
- Wilson (2006): highish μ's, weak constraints have high σ's and can thus be "demoted" easily.
- James White's oeuvre (cited last time): σ always the same; μ's directly reflect constraint strength.

11. Priors 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.
- Often the data "overwhelm the prior" (a favorite phrase of computer scientists).[2]
- You can mimic acquisition either by adding data (artificially with multiplication?)
  - Or you can increment the σ's; either way you change the balance of data to prior.

12. Making the μ's scientifically legitimate

- Ideally, they come from some legitimate source, not from wishful thinking to get the outcome right …

13. General notions of constraint strength I: OO-Correspondence

- Output-to-output correspondence is (at least sometimes) 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'.

  - Paper I can't divulge (under review) finds such effects, much more systematically, for verb paradigms in Korean.
  - The Jo paper suggests OO-correspondence for Tapping for English learning kids

[2] I spotted some thoughtful commentary on when overwhelming happens at https://stats.stackexchange.com/questions/200982/do-bayesian-priors-become-irrelevant-with-large-sample-size

14. General notions of constraint strength II: Markedness is stronger than Faithfulness (???)

- This goes back a long way, e.g.
  - Smolensky, Paul. 1996. The initial state and 'Richness of the Base' in Optimality Theory. Rutgers Optimality Archive ROA-154, http://ruccs.rutgers.edu/roa.html.
  - Hayes, Bruce (2004). Phonological acquisition in Optimality Theory: the early stages. In R. Kager, J. Pater & W. Zonneveld (eds.) Fixing priorities: constraints in phonological acquisition. Cambridge: Cambridge University Press. 158–203.
  - Prince, Alan and Bruce Tesar (2004). Learning phonotactic distributions. In R. Kager, J. Pater, and W. Zonneveld (eds.), Fixing priorities: constraints in phonological acquisition. Cambridge University Press, 245-291.
  - Boersma
- Subset Principle of learning: to make sure that bad things are classified as bad, you need to impose the restrictive system *a priori*, reversing it only under duress.
  - E.g., innate Coordinate Structure Constraint, unlearned only by Serbo-Croatians.
- I used to believe this (Hayes 2004).
  - Modern, probability-based methods of learning like maxent do not need to make such assumptions.
  - We may want to impose a priori strengths on our constraints to capture a learning bias.
  - But we do *not* need to do so to capture subset learning!

15. Following up: specifics on the modern solution to the "no negative evidence" problem

- We need to make a GEN that includes the bad cases we want to avoid.
- We also need a constraint set that makes the relevant structural distinctions.
- Maximum likelihood learning sucks up all the probability to the good cases.[3]
- This starves the bad cases of probability, even though nothing directly told us they are bad.

16. General notions of constraint strength III: Phonetically based priors

- Wilson, White both used **confusion matrix data** to derive measures of phonetic similarity, which then map onto priors for the OO-correspondence constraints.
- 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 the confusion rates.

17. A very toy example

- Inspired by (but executed rather differently from)
  - Jo, Jinyoung (2017) Learning Bias of Phonological Alternation in Children Learning English, M.A. Seoul National University.
- Data are somewhat thin; *just for pedagogy*; let's assume the following.

[3] As well as to non-heard cases that have the same constraint violations as the learning data.

  - Kids like Tapping less than adults do.
  - Tapping is easier for /d/ than /t/.
  - They often produce [d] as the output for tappable /t/ (I have noticed this myself in observing children.)
- [to spreadsheet]

WILSON AND OBDEYN (2009) ON MODEL EVALUATION: A READER'S GUIDE

18. Why we are reading it

- It is part of a long series of papers on an intriguing aspect of phonology; the **avoidance of similar consonants in roots**.
  - Wilson and Obdeyn think they have arrived at the right theory now.
- Forceful advocacy of models based on the theory of probability, vs. ad hoc metrics like Observed/Expected.
  - This is a pervasive point made in Jaynes's probability book, cited in Handout 1.
- Review of maxent (description from another expository tradition)
  - I have noticed it is a skill to recognize the same thing described differently by different technical people.
- Emphasis on finding the right theory as a **blend of accuracy and restrictiveness**.

19. Recommendations for reading it

- Make your way through, absorbing the main points.
- If you get stuck, don't obsess, but move on, relying on the outline here for what you might want to get on first reading.
  - I freaked out on my first reading, but am getting more comfortable after my fourth. I still don't understand the Laplace approximation.

20. The basic phenomenon

- Avoidance (absolute or relative) of similar consonants in stems, like /bapal/ or /ralak/ or /tap'ap/.
- This is found all over the world: Arabic, Muna, English, Wargamay, Shona, Quechua …
- Wilson/Obdeyn seem to think you will find it wherever you look for it and I know of no counterexample.

21. The special status of coronals

- All labials tend to not want to occur with any labials.
- All dorsals tend to not want to occur with any dorsals.
- Coronals cooccur more, and dislike cooccuring with coronals similar in other features like manner.

22. Psychological reality of such effects

- A nice experiment, using non-college subjects:

  - Frisch, S. A. and Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, 77:91–106.
- "OCP-Place" means Obligatory Contour Principle; don't have two C of same place.

23. Ur-papers first treating the phenomenon

- McCarthy, J. J. (1988). Feature geometry and dependency: a review. Phonetica, 45, 84–108.
- Pierrehumbert, Janet B. (1993). Dissimilarity in the Arabic verbal roots. In Schafer, A., editor, *Proceedings of the North East Linguistic Society* 23, pages 367–381, Amherst, MA. GLSA.
- Pierrehumbert also introduced (I think) the **observed over expected** method for assessing underrepresentation and overrepresentation.

24. A characteristic pattern observed by McCarthy: his constraints

- OCP-LAB Adjacent labials are prohibited
- OCP-DOR Adjacent dorsals are prohibited
- OCP-PHAR Adjacent pharyngeals are prohibited
- OCP-COR[αSON] Adjacent coronals agreeing in [+/–sonorant] are prohibited

- So coronals combine more freely, but their subsets defined by manner are still regulated.

THE OBSERVED/EXPECTED METHOD

25. Step-by-step calculation (for consonant cooccurrence problems)

- Form a *n* by *n* chart of consonants and count all … C_1 V C_2 … for each cell. For Wilson and Obdeyn's contrived data:[4]

	P_2	T_2	K_2
P_1	345	1034	345
T_1	1034	776	517
K_1	345	517	86

- Total the rows and columns, and indeed the whole chart:

		P_2	T_2	K_2	all
		1724	*2327*	*948*	*4999*
P_1	*1724*	345	1034	345	
T_1	*2327*	1034	776	517	
K_1	*948*	345	517	86	

[4] Contrived to assassinate the approach!

- Calculate fraction of total consonants in the relevant position for both C1 and C2. Here they are identical because these are contrived data.

		P_2	T_2	K_2
		0.345	0.465	0.190
P_1	0.345			
T_1	0.465			
K_1	0.190			

- Multiply the values out to get expected proportions of individual cells.

		P_2	T_2	K_2
		0.345	0.465	0.190
P_1	0.345	0.119	0.161	0.065
T_1	0.465	0.161	0.217	0.088
K_1	0.190	0.065	0.088	0.036

- Multiply these values by the total number of consonants to get the Expected values:

	P_2	T_2	K_2
P_1	594.6	802.5	326.9
T_1	802.5	1083.2	441.3
K_1	326.9	441.3	179.8

- Divide the Observed cells (starting point) by Expected cells to get a number, O/E:

	P_2	T_2	K_2
P_1	0.58	1.29	1.06
T_1	1.29	0.72	1.17
K_1	1.06	1.17	0.48

- This we can really understand: PP, TT, KK are underrepresented, and the noncoronals are particularly bad.
- This conclusion turns out to be utterly bogus! This is the point of Wilson and Obdeyn's example.

26. Back to the data: The curious override to the similarity-avoidance principle

- Arabic is a canonical case of avoiding similar consonants in roots, but it also loves "biliteral" roots with *identical* consonants.
  - /samam/ 'to poison'
  - This was an elegant focus of John McCarthy's 1979 Ph.D. dissertation, where he used autosegmentalism to make such roots genuinely bilateral.

- /samam/ is ok because of left-to-right "spreading", */sasam/ is not ok.
- Autosegmental morphology is mostly gone, I think, due to contradictions: sometimes you want consonants on a separate tier, sometimes the same.

27. Modern views of the override

- Zuraw (2002) has found that speakers love to interpret roots as *reduplicated* if they can, and exaggerate the degree of resemblance through mispronunciation: [pɑmpɑm] for *pompon.*
  - Zuraw, Kie (2002). Aggressive reduplication. Phonology 19. Pp. 395-439.
- So /samam/ looks perhaps like constraint ranking; it's horrible for similarity, great for "be reduplicated".
- Coetzee and Pater (2008) later noted, re. the same sort of override in Muna:
  - "In particular, in many cases the identical consonants do precede identical vowels, suggesting some form of reduplication"

28. A spectacular proposal

- Early version:
  - Frisch, S. A., Broe, M. B., and Pierrehumbert, J. B. (1997). Similarity and phonotactics in Arabic. Bloomington, IN and Evanston, IL: Indiana University and Northwestern University, ROA-223.
- Later:
  - Frisch, S. A., Pierrehumbert, J. B., and Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, 22(1):179–228.

- Key idea:
  - There is *one* mechanism for computing similarity in phonology, and you don't even have to do phonetics to access it!
  - You need to make a list of all the natural classes in the language (often about 600, easy to do with computer)
  - Then, for two segments, compute

    (shared natural classes) / ((shared natural classes) + (unshared natural classes))

  - Dissimilarity is then predicted by taking this one single dissimilarity measure as a constraint.
- This amazingly simple and restrictive theory got quite a lot of empirical mileage!
- It got the special status of coronals for free: there are *more coronals* in virtually any language, and so more natural classes involving them, and so in general less similarity between them.
- People like me adopted the similarity metric for other purposes (e.g. Albright and Hayes 2001, *Cognition*).

29. Coetzee and Pater's contributions

- Coetzee, A. W. and Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory*, 26:289–337.
- Language specificity: What works for Arabic doesn't work for Muna (Western Austronesian)
  - "Arabic shows an overwhelming effect of sonorancy agreement in lowering attestedness, while Muna has a more balanced contribution of voicing, sonorancy and stricture."
- We are in familiar territory: learning phonotactics with multiple constraints, using a data corpus.
- They also advocate a particular measure of phonotactic well-formedness:
  - "We argue for a definition of Harmony-based well-formedness in terms of the difference between the [harmony] score of an Input-Output mapping and the optimal distinct mapping for the same Input."
  - For instance, for [t … s] the best output candidate might be [p … s].
- Constraints: they advocate a *very rich* model, in which constraints target specific combinations of place feature and manner feature.

  "Assign a violation mark to a sequence of nonidentical consonants that both have place of articulation P and agree in specification for S; where
  P is drawn from the set {Pharyngeal; Dorsal; Coronal; Labial}
  S is drawn from the set {Sonorancy; Stricture; Voice; Emphatic; Prenasalization}"

- Summarizing:
  - Discovery of language-specific effects
  - Modeling of languages with inventories of phonotactic constraints and a Harmony-based framework.

WILSON AND OBDEYN'S TRASHING OF OBSERVED/EXPECTED
(AND PAEAN TO MAXENT)

30. The data covered in demo (25) are artificial

31. The origin of their numbers: Wilson/Obdeyn's calculations

- They sum to 5000 (rounding aside).
- The proportions for rows are 1/3, 1/2, 1/6 for P T K.
- Ditto for columns.
- Lastly, they halved the values on the diagonal.
- The cells just reflect multiplication of these basic frequency principles.

32. The fatal artifact

- Plainly, the system is "designed" with an equal effect of OCP, across all three places.
- Yet O/E statistics ((24) above) tells us there is a *stronger* OCP for labials and dorsals!

33. Let's check on our own: do a classical maxent analysis of the same data[5]

- Reformat the data in rows:

p	p	345
p	t	1034
p	k	345
t	p	1034
t	t	776
t	k	517
k	p	345
k	t	517
k	k	86

- Here are baseline constraints:[6]
  - *[p] as C1
  - * [t] as C1
  - * [k] as C1
  - * [p] as C2
  - * [t] as C2
  - * [k] as C2
  - *Classical* OCP (no identical C in CVC)

- The grammar is, unsurprisingly a perfect fit:

			C1 p	C1 t	C1 k	C2 p	C2 t	C2 k	OCP	H	eHar-mony	Z	P	obs
			0.90	0.50	1.60	0.90	0.50	1.60	0.69					
p	p	345	1			1			1	2.499	0.082	1.191	**0.069**	**0.069**
p	t	1034	1				1			1.401	0.246	1.191	**0.207**	**0.207**
p	k	345	1					1		2.499	0.082	1.191	**0.069**	**0.069**
t	p	1034		1		1				1.401	0.246	1.191	**0.207**	**0.207**
t	t	776		1			1		1	1.688	0.185	1.191	**0.155**	**0.155**
t	k	517		1				1		2.094	0.123	1.191	**0.103**	**0.103**
k	p	345			1	1				2.499	0.082	1.191	**0.069**	**0.069**
k	t	517			1		1			2.094	0.123	1.191	**0.103**	**0.103**
k	k	86			1			1	1	3.885	0.021	1.191	**0.017**	**0.017**

[5] This is a common practice of statisticians: try your methods on data you concocted yourself, so that the causal mechanisms you want to detect are actually known.

[6] They all turn out to be constraints, not virtues, with positive weights.

- Socratic questions
  - I later added three new constraints: OCP(labial), OCP(coronal), OCP(dorsal).
  - What weights are they likely to receive?
  - What would be their performance on the likelihood ratio test?

34. Can we detect the intentions of the designer in the weights themselves?

- Yes!
- Use a simple theorem:
  - In maxent, the difference in harmony between two candidates is the **log of the odds** of the two candidates.
  - where "odds" = ratio of probabilities
- Imagine the candidate [par], which violates $*P_1$ once.
- Imagine the candidate [tar], which violates $*T_1$ once.
- Their harmonies would be just the weights of these single constraints.
  - [par]: 0.903459744
  - [tar]: 0.498306781
- Difference: -0.405152963
- Undo log: $e^{-0.405152963}$ is 0.666874796, or two thirds
- This is exactly the ratio used in designing the data set (see (30): one third against one half).
- Similarly, a candidate like [rar], with just OCP violation, would get half the probability of a violation-free candidate like [ral].

35. Upshot of the example

- Maxent (with suitable constraint choices by us) correctly detected the intentions of the designer in these concocted data: no special treatment for coronals etc.
- O/E statistic is grossly misleading on this point, serving up an artifact.

36. Or more abstractly

- We seek to understand the statistical properties of a corpus — essentially probabilities.
- So best to set up mechanisms (constraints), use them in a rational, mathematically well-motivated way to model probabilities; success is support for the constraints.
- Superficial metrics of goodness are unreliable — because the computation of actual probabilities is a delicate interaction of multiple factors.

MODEL EVALUATION IN WILSON AND OBDEYN

37. The primary method

- Likelihood, as we have been working with.
- This is the metric of *accuracy*.

38. Accuracy is not enough

- Model complexity must also be determined.
- Simple example proving this: a model with these constraints:
  - *p p, *p t, *p k, *t p, *t t, *t k, *k p, *k t, *k k (and so on for all possible pairs)

  cannot but succeed in provide a *perfect match to data*, but is worthless as an explanation.

39. Where they head with the math

- Laplacean approximation, which combines a measure of accuracy with a measure of model restrictiveness.

BY THE WAY, WHO WON?

40. The narrow question at hand

- Coetzee and Pater's claim to have improved on Pierrehumbert's team is refuted; it's more of a tossup.
- Wilson and Obdeyn's theory:
  - maxent and probabilities, of course
  - Constraints are a simplification of Coetzee and Pater's:

    normal *Lab Lab, *Cor Cor, etc.
    Role of manner:
    "Don't agree in [voice], [son], etc., *if you agree in place*"
    (so, simpler; not crossing every place and manner feature)
    (the italicized bit perturbs me and I'd like to know how well we can do without it)
- The clear winner by the Laplace approximation criterion is the theory (simplified Pater-Coetzee; manner features treated the same everywhere) of Wilson-Obdeyn.
- I would love to know what happens if you do something even simpler: just weight all the features, punkt.

41. The explanation for indifference of labials and velars to manner

- It's really trivial, if Wilson/Obdeyn are right:
  - Saturation
  - Weights of OCP-lab and OCP-dorsal are high enough that further nuances from manner don't affect probability much, (remember the purposes of exponentiation in maxent).