Class 5, 4/14/2020: Last bit on frameworks, O/E

1. Assignments

- Read:
  - If you have time start reading: Lise Menn (1983) Development of articulatory, phonetic, and phonological capabilities. In Brian Butterworth, *Language Production* vol. 2. Academic Press.
  - You can finish reading it next time.
  - On course web site
- Finish your homework on medial clusters.
  - Due 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

- UG bias and how to implement it in grammar learning — Moore-Cantwell example
- The debates over MaxEnt and NHG: harmonically bounded winners, overgeneration
- The Observed/Expected fallacy and how to avoid it in MaxEnt

3. Coming up

- Phonological acquisition — empirical, then a bit of theoretical

4. Review of bias so far

- The grand scheme:
  - a UG-guided learning model for all of language.
  - The UG being, perhaps, soft (bias, not absolute)
  - Many, many ways we might imagine this to happen
- The MaxEnt way, invented by Wilson
  - The Universal Constraint Set of OT is assigned a priori weights (based on some collection of theories)
  - Learning takes place as these a priori weights are modified by confrontation with data.

5. A toy example of bias: 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. [☞ Or last time: inattention; variable participation, different lexicon from what we used …]
  - We model this with the same prior, mean zero, sigma to be determined, for all constraints.
- Is it legit to study such a tiny case?
  - We have four data points, four parameters, ouch.
  - Yet the theory cannot compute any set of values, so there is in principle some sort of legitimacy to this.
- Try the spreadsheet

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

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

7. Calculating the needed bias from experimental data

- The White method:
  - Find a confusion matrix (phoneme *x* heard as *y*)
  - Fit a MaxEnt model *CONFUSE *x* WITH *y*
  - Let the weights be the prior weights for *MAP(*x*, *y*)

SURVEY OF THE FRAMEWORK-EVALUATION LITERATURE

8. Bashing of Stochastic OT

- You've read the Zuraw/Hayes bashing based on wug-shaped curves — related to ignoring of data, common sense judgment on arriving at conclusions.

9. Bashing of Noisy Harmonic Grammar

- In the form naturally favored by the research community (?), it assigns zero probability to harmonically bounded candidates.

10. Quick review of Harmonic Bounding

- ☞ Tell me the harmonically bounded candidates in this tableau

			Ident(voice)	Ident(sonorant)	Don't not tap
pat	pa[t]	1			
	pa[d]		1		
	pa[ɾ]		1	1	
patting	pa[t]ing	0.2			1
	pa[d]ing		1		1
	pa[ɾ]ing	0.8	1	1	
pad	pa[t]		1		
	pa[d]	1			
	pa[ɾ]			1	
padding	pa[t]ing		1		1
	pa[d]ing	0.1			1
	pa[ɾ]ing	0.9		1	

11. The details of excluding harmonically bounded candidates in NHG

- Reasoning:
  - Let A have a subset of the violations of B.
  - Noise is assigned at the *constraint* level.
  - Now multiply the invariant perturbed noise by asterisks, and compute harmony.
  - If the perturbed weights are positive, the computation will give B a greater aggregate Harmony penalty than A, and A will win.
- ☞ Tell me how this would work to exclude *[paɾ]
- Upshot: if we need to assign probability to harmonically bounded candidates, then this version of Noisy Harmonic Grammar cannot be right.

12. Worm in the apple

- How to keep *perturbed* weight positive?? See Boersma and Pater (2016) for one method, using log weights. This is little explored.
- My own software (OTSoft) doesn't yet use log weights and I find that often harmonically bounded candidates *do* win.

AN UNDEREXPLORED AREA WHERE HARMONICALLY BOUNDED CANDIDATES MIGHT NEED TO RECEIVE PROBABILITY: PHONOLOGICAL PHRASING

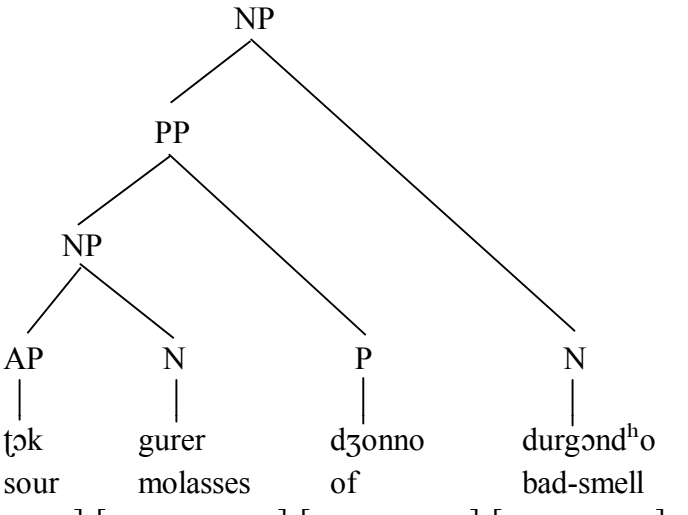
13. A bit on phonological phrasing

- Long ago both Hayes/Lahiri and Jun noticed "unmotivated" free variation in the formation of phonological (a.k.a. accentual) phrases in Bengali resp. Korean.

- Bruce Hayes and Aditi Lahiri (1991) Bengali intonational phonology. *Natural Language and Linguistic Theory* 9: 47-96.
- Jun, Sun-Ah (1996) *The phonetics and phonology of Korean prosody: intonational phonology and prosodic structure*, Garland Publishing Inc., New York : NY

- For Bengali, there are abundant diagnostics (intonational tones, segment assimilations) that tell you the correct phrasing; true for Korean too.
- Bengali: to understand the system, it mostly suffices to look at long left- and right-branching structures.

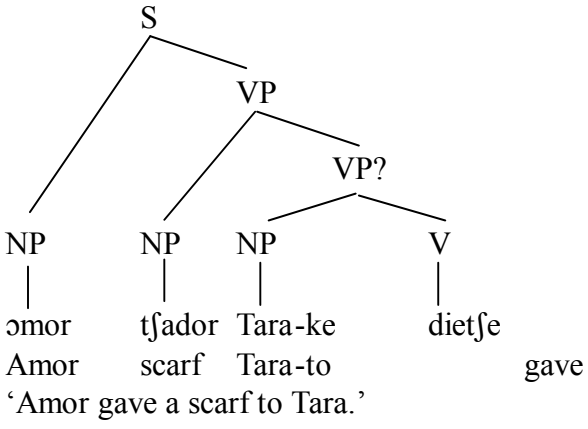
14. Left branching



[	]P[	]P[	]P[	]P
[		]P[	]P[	]P
[			]P[	]P
[				]P
*[	]P[	]P[		]P
* [	]P[		]P[	]P
* [		]P[		]P
* [	]P[			]P

- Don't group two right branches into a single phrase.
  - Korean: same, only left!

15. Right branching



- Phrase verbs separately (see Hayes/Lahiri for a rationale: keeps phrasal verbs distinct from the numerous compounds).

16. A big tie when we put in only the inviolable constraints

Input	Candidate	Target	Pre-dicted	ALIGN V	*2 RIGHTS
O to S sari gave	* [Orundhoti][Shamoli][scarf gave]	0	0	*	
	* [Orundhoti][Shamoli scarf gave]	0	0	*	
	* [Orundhoti Shamoli][scarf gave]	0	0	*	
	* [Orundhoti Shamoli scarf gave]	0	0	*	
	[Orundhoti][Shamoli][scarf][gave]	0.25	0.25		
	[Orundhoti Shamoli][scarf][gave]	0.25	0.25		
	[Orundhoti][Shamoli scarf][gave]	0.25	0.25		
	[Orundhoti Shamoli scarf][gave]	0.25	0.25		
sour molasses for stink	*[sour][molasses for][stink]	0	0		*
	* [sour][molasses][for stink]	0	0		*
	* [sour molasses][for stink]	0	0		*
	* [sour][molasses for stink]	0	0		**
	[sour][molasses][for][stink]	0.25	0.25		
	[sour molasses][for][stink]	0.25	0.25		
	[sour molasses for][stink]	0.25	0.25		
	[sour molasses for stink]	0.25	0.25		

17. Extending the system to accommodate speech rate

- Perhaps simply *PHRASE, a version of *STRUC?
- This gets bigger, the faster/less carefully a Bengali speaks.
- Perhaps it is negative in the slower speaking rates.
- Try this out on the spreadsheet.
- Ponder also the harmonic bounding situation.

18. The key issue for frameworks respecting harmonic bounding

- To find *virtues* for every specific winner.[1]
- This arises especially for the **multiple-locus problem** in phonology, studied by Bert Vaux, Aaron Kaplan, and others.

19. The other argument for letting harmonically-bounded candidates get probability

- The GEN-based theory of phonotactic well-formedness, which you are working with right now on the homework.

BASHING MAXENT (ALSO NHG)

20. Maxent overgenerates

- Arto Anttila, Scott Borgeson, and Giorgio Magri (2019) Equiprobable mappings in weighted constraint grammars. Posted on arXiv.
  - Scary result for Finnish stress; I've asked Anttila for data so I can check myself.
- Anna Mai and Eric Bakovic (2019) Cumulative constraint interaction and the equalizer of HG and OT. AMP2019

21. A possibly scary possible overgeneration from Bakovic and Mai

- They demo a system in which
  - /sa/ → [sa]
  - /ʃa/ → [sa]
  - /si/ → [si]
  - /ʃi/ → [si]
- Let's give it a try in MaxEnt.
- Since the problem is Harmonic Grammar itself, NHG also can derive this system.
- Stochastic OT cannot.
- Note that this is an unattestedness argument.
  - accidentally unattested

[1] A very interesting effort is Hubert Truckenbrodt (2002) Variation in p-phrasing in Bengali. Linguistic Variation Yearbook 2 (2002), 259–303. He uses lots of OO correspondence, essentially cyclicity at the phrasal level.

  - unattested for historical reasons?
- Recent work on unattestedness arguments
  - The expressivity of segmental phonology and the definition of
  - weak determinism, Adam G. McCollum, Eric Bakovi¢, Anna Mai, Eric Meinhardt
  - Hayes and Jo on Balinese, Hayes web site

22. Is MaxEnt really tested enough?

- The lazy bit
  - Every non-stochastic analysis in Classical OT has a MaxEnt near-equivalent.
  - You just have to set the weights high enough to approximate strict ranking.
  - For discussion (but on non-stochastic HG), see Prince, Alan. 2002. Anything Goes. In *A New Century of Phonology and Phonological Theory*, ed. T. Honma, M. Okazaki, T.Tabata, & S. Tanaka , also Rutgers Optimality Archive.
- The need for vigilant checking: when you do analysis in maxent, you must include harmonically bounded candidates in the candidate set.
  - We lose a luxury that we had in classical OT analysis.

23. Demo of Danger, Harmonic Bounding

- Let us be happy naïve Maxent users, and work on this tableau:

			Ident (voice)	Ident (son)	Don't not tap
patting	pa[t]ing	0.2			1
	pa[d]ing		1		1
	pa[D]ing	0.8	1	1	
padding	pa[t]ing		1		1
	pa[d]ing	0.1			1
	pa[D]ing	0.9		1	

- We will succeed, more or less.
- But we forget to check the logical possibility of *spontaneous* tapping or voicing changes!

			Ident (voice)	Ident (son)	Don't not tap
pat	pa[t]	1			
	pa[d]		1		
	pa[D]		1	1	
patting	pa[t]ing	0.2			1
	pa[d]ing		1		1
	pa[D]ing	0.8	1	1	
pad	pa[t]		1		
	pa[d]	1			
	pa[D]			1	

padding	pa[t]ing		1		1
	pa[d]ing	0.1			1
	pa[D]ing	0.9		1	

- This turns out to be problematic.
- IDENT(voice) cannot be weighted too high, since it serve as a mere "modulator" for the frequencies of tapping (italic frequencies).
- But then we get spontaneous changes of voicing in plain stems.

24. The maligned Stochastic OT did fine on this one

- Version run: in my own OTSoft software.

```
/pat/      Input Fr. Gen Fr.  Input #     Gen. #
pa[t]        1.000   1.000      24605     100000
pa[d]        0.000   0.000
pa[D]        0.000   0.000

/patting/  Input Fr. Gen Fr.  Input #     Gen. #
pa[t]ing     0.200   0.201       4980      20059
pa[d]ing     0.000   0.000
pa[D]ing     0.800   0.799      19983      79941

/pad/      Input Fr. Gen Fr.  Input #     Gen. #
pa[t]        0.000   0.000
pa[d]        1.000   1.000      25240     100000
pa[D]        0.000   0.000

/padding/  Input Fr. Gen Fr.  Input #     Gen. #
pa[t]ing     0.000   0.000
pa[d]ing     0.100   0.103       2523      10345
pa[D]ing     0.900   0.897      22669      89655
```

25. Diagnosing the success of Stochastic OT

- Ident(voice) looks weak in its capacity as a modulator of Tapping.
- But it *has no competition* in its capacity as enforcer of phonemicness, and so who cares if it is weak.

26. The eternal response to problems in constraint-based theories

- Maybe the constraints are not right, or there are problems with the grammatical architecture.
- E.g. can we profitably regard Ident(voice) as a strong principle of phonotactics, but weak in alternations?

THE OBSERVED/EXPECTED FALLACY AND HOW TO AVOID IT IN MAXEN T

27. Observed/Expected

- This is a very widely used method of assessing phonotactic data, developed (?) by Pierrehumbert (1993).
- It is defective and should not be used.
- For user-friendly discussion, see Appendix of Breiss/Hayes (forthcoming in Lg.)

28. Where I learned about this

- Colin Wilson and Marieke Obdeyn (2009) Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms., Johns Hopkins University. Easily Googled.

29. Calculating Observed/Expected for consonant cooccurrence problems

- These could be clusters, or perhaps non-local dependencies as in Arabic.
- 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 pseudo-Arabic data:

	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	

- 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.

30. The origin of these 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.

31. The fatal artifact

- Plainly, the system is "designed" with an equal effect of OCP, across all three places.
- Yet O/E statistics ((27) above) tells us there is a *stronger* OCP for labials and dorsals!

32. Let's check on our own: do a classical maxent analysis of the same data[2]

- 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:[3]
  - *[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**

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

- Yes!
- Use the principle taught earlier:

[2] 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.

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

  - 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].

34. 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.
- See Breiss/Hayes for a rather different example making the same point.

35. 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.