Linguistics 219 Phonological Theory III
B. Hayes Spring 2019

Class 3, 3/9/2018: More on Frameworks; Bias I

1. Assignments

- Read:
  - Michael Becker, Andrew Nevins, and Jonathan Levine (2012) Asymmetries in generalizing alternations to and from initial syllables. *Language* 88:2, pp. 231–268.
  - An ardently UG-ist paper on universal biases.
- Continue homework on medial clusters.
  - Due in a week, Monday April 16.

A BIT MORE ON WARLPIRI CONSONANT CLUSTERS

2. The n by n chart

	p	t	ʈ	c	k	m	n	ɳ	ɲ	ŋ	l	ɭ	ʎ	ɾ	ɽ	ʂ	ɻ	w	j
p																			
t	3			2	1														
ʈ	1			1															
c	3																		
k																			
m	113																		
n	92	11			61	14				16									
ɳ	76		142	8	33	8				17									
ɲ	53				15	1													
ŋ					132														
l	14	7		52	77	1												5	
ɭ	85		14	16	6					2								1	
ʎ	89			2	33					2									
ɾ	171			26	117	22				18								8	1
ʈ																			
ɻ	4				1	1			1										
w																			
j	1																		

- I found this very useful.
- Note that columns and rows are sorted in strict IPA order.

3. Where I am currently

- Best analysis is an augmentation of Nash's account.

- Likelihood of the data = -6526.8, a bit better than my freshly-created -6594.7

4. Where to stop

- This seems to be something of an art.
- It seems ill-advised to include constraints in the completed model that don't test significant.
- Hayes, Wilson, and Shisko (2012, *LI*), armed with 87 constraints they wanted to test, tried:
  - bottom up (add best constraint till no significant improvement)
  - top down (delete worst constraint until you would delete a significant one)

  with similar results.

FINISHING OUR TOUR OF THE FRAMEWORK BAZAAR

5. 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.

6. Many varieties exist

- See
  - Bruce Hayes (2017) Varieties of Noisy Harmonic Grammar. *Proceedings of the 2016 Annual Meeting in Phonology*, USC.
- E.g., where do you put the noise?
  - On the constraint weights (= classical version)
  - In the tableau cells
  - On the harmony values (behaves amazingly like maxent)
- If there are several violations of a constraint, does this cause the noise to be similarly multiplied?

7. Assessment

- I personally feel this framework is in contention:

- Performs about as well in practice (I suspect) as maxent.
- No proof of convergence for learning algorithm, but I have never seen it misbehave.
- Combines evidence from multiple sources in making predictions (in the very same way as maxent, its partner in stochastic Harmonic Grammar).

RETURN FROM THE BAZAAR TO PONDER: WHAT ARE THE ISSUES IN FRAMEWORK CHOICE?

8. Ganging

- I've portrayed ganging as deeply rational and wholesome, but what are the linguistic facts?
- Perhaps one could say that ganging is increasingly noticed in phonology as people look for it.
- Some feel (unpublished work of Edward Flemming) that ganging is a property of optional phonology, and that "crystallized", obligatory phonology doesn't gang.
  - A tall order to explain, and worth pondering.

9. Harmonic bounding

- A harmonically bounded candidate in OT has a strict superset of the violations of a rival candidate.
- In classical OT, it can never win.
- In stochastic OT, it can never win.
- In maxent, it can, but never with the highest probability.
  - Mechanism: superset of constraints, so higher harmony than bounder, so lower eHarmony than bounder, so lower probability than bounder.[1]
- Noisy Harmonic Grammar: usually it can (see Hayes paper), but there is one little-explored variant (Exponential Noisy Harmonic Grammar; Boersma/Pater 2017), in which it cannot.

10. Harmonic-bounding Implication I

- Be very careful when you do analysis in maxent, because you must include harmonically bounded candidates in the candidate set.
  - We lose a luxury that we had in classical OT analysis.
- In theories with no harmonic bounded winners, we can often find all the reasonable candidates by asking "how else could the fatal Markedness constraint be repaired?"
- In maxent, we are better off just finding all the combinatorial possibilities of relevant factors and listing their free combinations.

[1] This assumes weights greater than zero for the relevant constraints. If we are going to allow negative weights, than the definition of "harmonically bounded" has to be adjusted accordingly.

11. Demo of Danger, Harmonic Bounding

- We return to the Class 1 demo of tapping, and now include tapped candidates for the plain stems *pat* and *pad*.
- These can harmlessly repaired, but we need to put in further constraints on the distribution of tap.
- See spreadsheet for what happened.

			Ident (voice)	Ident (son)	Don't not tap			
			3.3	**12.2**	**16.4**	Harmony	p	observed
pat	pa[t]	1				0.000	0.963	1.000
	pa[d]		1			3.261	**0.037**	
	pa[D]		1	1		15.443		
patting	pa[t]ing	0.2			1	16.401	0.274	0.200
	pa[d]ing		1		1	19.661	0.011	
	pa[D]ing	0.8	1	1		15.443	*0.715*	0.800
pad	pa[t]		1			3.261	**0.037**	
	pa[d]	1				0.000	0.963	1.000
	pa[D]			1		12.182		
padding	pa[t]ing		1		1	19.661	0.001	
	pa[d]ing	0.1			1	16.401	0.014	0.100
	pa[D]ing	0.9		1		12.182	0.985	0.900

- 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 (bold).

12. The maligned Stochastic OT did fine on this one

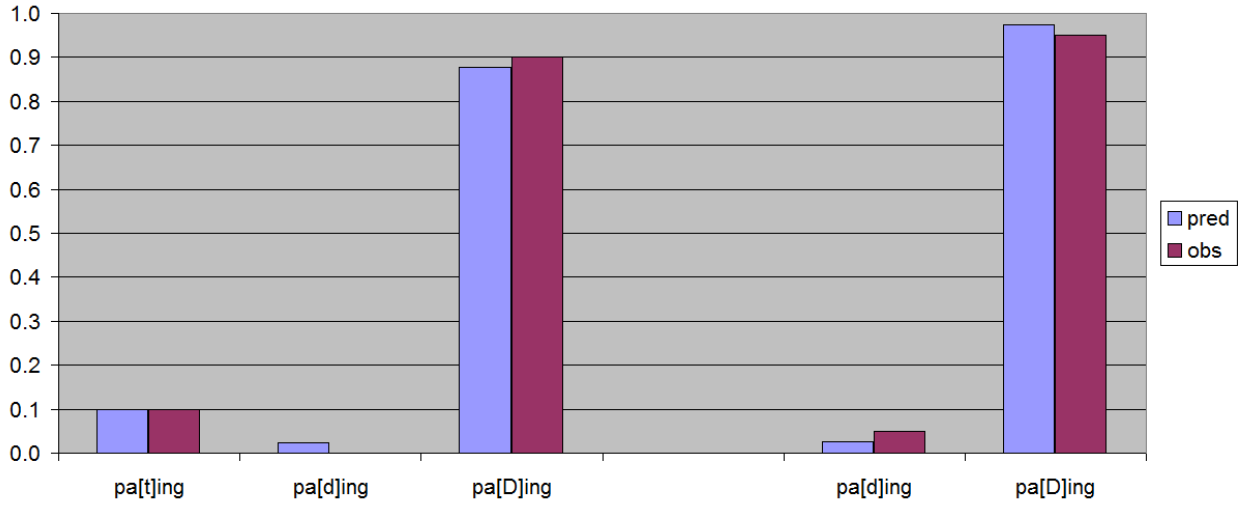
- Version run: in my own OTSoft software.

```
/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
```

13. A somewhat-repaired maxent version of the Tapping problem

			Id(voice)-final	*Map(+ -voice)	*Map(- +voice)	*Map(- +son)	Don't not tap
			50.00	50.00	1.44	0.50	4.11
pat	pa[t]	1					
	pa[d]		1				
	pa[D]		1				
patting	pa[t]ing	0.1					1
	pa[d]ing				1		1
	pa[D]ing	0.9			1	1	
pad	pa[t]		1				
	pa[d]	1					
	pa[D]		1				
padding	pa[t]ing			1			1
	pa[d]ing	0.05					1
	pa[D]ing	0.95				1	



- *Map(x, y) is meant to regulate paradigms; see
  - Zuraw, Kie (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. *Language* 83. Pp. 277-316.
  - Zuraw, Kie (2013). *MAP constraints. Unpublished manuscript, on web site.
- My *MAP constraints are directional; which is a move proposed at various places in the OT literature (ran out of time to find this! search on Max(feature).
- Note the modest probability allocated to /pætɪŋ/ → [pæ**d**ɪŋ]. This might actually be correct; see below on Bias.

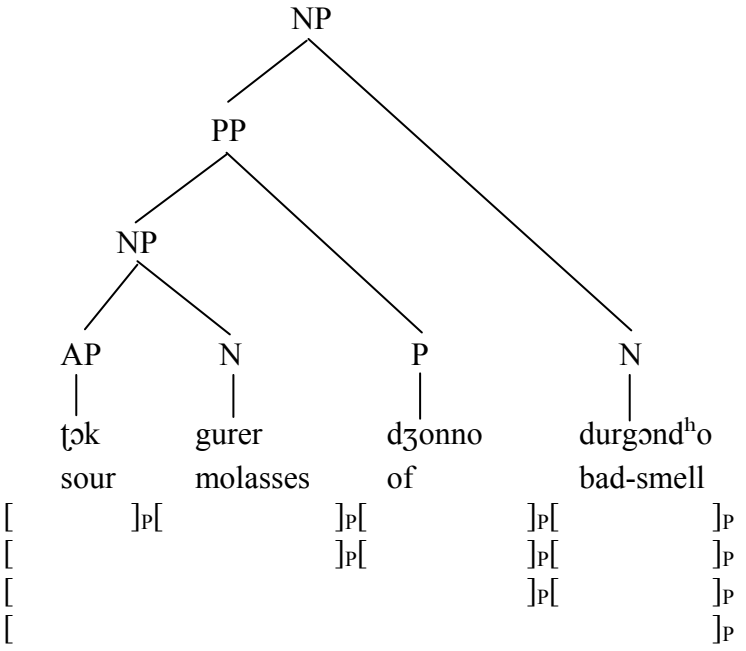
14. Harmonic-bounding Implication II

- It becomes empirically important whether harmonically bounded candidates win in real life.
- I think they can be found in:
  - Phonotactics (the Markedness-only approach we are working with now)
  - Metrics (see Hayes and Moore-Cantwell 2012 *Phonology*, paper with Russ Schuh under revision)
  - Syntax-phonology interface: multiple phrasings from one syntactic structure.
- Harmonic bounding currently has Mom-and-Apple-Pie status (restrictiveness, ease of analysis) in phonology, and it will take a lot of empirical argument for it to lose this status.

15. A place to look for harmonically bounded part-winners: 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.

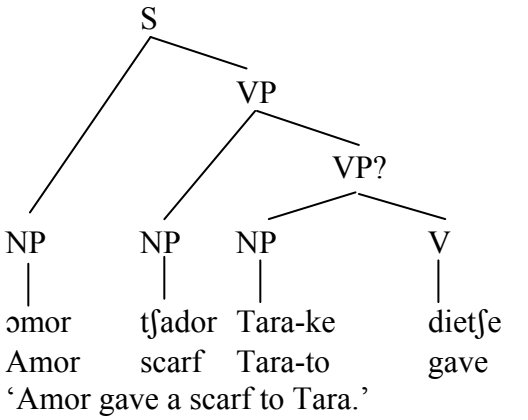
16. Left branching



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

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

17. Right branching



'Amor gave a scarf to Tara.'

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

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

18. This all swims nicely in maxent

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		

- I see it, perhaps, as a dilemma for other theories to find *virtues* for every specific winner.[2]
- We might proceed further to try to model frequencies more precisely; generally the phrases get bigger as one speaks faster.

BIAS

19. Soft UG

- I had no idea when I was told about Linguistics Universals as a teenager how many proposed "hard" universals would bit the dust!
- Some recent favorites:
  - Serbo-Croatian can extract wh-words from coordinate structures. (Daniela Culinovic, personal communication).
  - Vowel harmony processes can look ahead several syllables for opaque vowels, deciding therefore not to apply at all ("sour grapes")[3]
  - Warlpiri vowels take a vote on whether to be all front or all back, when harmony-irregular words get regularized (Margit Bowler, ms.).
- But experimental work is nonetheless discovering some tentative evidence for UG
  - **non-veridical learning**

[2] 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.

[3] McCollum, Adam G. & James Essegbey. 2017. Vowel harmony is not always myopic: Evidence from Tutrugbu. Proceedings of *WCCFL* 35.

- either in the experimental participant's life experience, or in an artificial language learning experiment

20. How to obtain 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[4] don't undergo.

21. 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 principles 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 Elliott Moreton (2008) Analytic bias and phonological typology; phonology 2008
- The readings for this time: 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.

[4] Except BH's *epitaph*

22. Toward modeling such effects: the Gaussian prior

- 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 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

23. 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.
- 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 one are powerful; sigmas like 100000 are virtually absent.
- I think I remember why it's called Gaussian: if you want to convert the expression above to true probability (not log probability) you must take e to the whole thing, which produces a Gaussian curve:

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

24. 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.

25. A 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.
- 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.

26. Determining the prior in modeling work

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

27. 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 adding (artificially?) data, or shrinking sigmas.

28. Making the mu's rigorous

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

29. 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'.

- Markedness is strong than Faithfulness
  - Because of the Subset Principle of learning: to make sure that bad things are classified as bad, you need to impose the restrictive ranking a priori.
  - I used to believe this; modern methods of learning like maxent do not need to make such assumptions however.

30. 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.

31. 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 quite thin; *just for pedagogy*; let's assume the following. Caveat: probably quite wrong!
  - 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.)