Class 9: More about MaxEnt; lexical variation; grammar architecture

To do for tomorrow
- Prepare your presentation and handout.

Overview: A bit more about MaxEnt and what kinds of patterns it can capture. Lexical variation. Variation and the architecture of the grammar.

1 Straggling point from last time

- I forgot to explain last time why the "prior", or "smoothing term" below is called a "Gaussian prior":
  - Maximize: ln(*probability*(data under model)) - $\sum_{j=1}^{M}\frac{(w_j-\mu_j)^2}{2\sigma_j^2}$
- The equation for the normal distribution, also know as Gaussian distribution, is $y=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ (I'll illustrate this on the board)
- If we want to maximize: $\ln(prob(data))+\sum_{j=1}^{M}\ln\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{\frac{-(w_j-\mu_j)^2}{2\sigma_j^2}}\right)$, that's equivalent to
- maximize: $\ln(prob(data))+\sum_{j=1}^{M}\left(\ln\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}\right)+\ln\left(e^{\frac{-(w_j-\mu_j)^2}{2\sigma_j^2}}\right)\right)$ =

  $\ln(prob(data))+a\_number\_that\_doesnt\_depent\_on\_weights+\sum_{j=1}^{M}\frac{-(w_j-\mu_j)^2}{2\sigma_j^2}$
- only thing learner can change is weights, so same as maximizing $\ln(prob(data))-\sum_{j=1}^{M}\frac{(w_j-\mu_j)^2}{2\sigma_j^2}$

2 Harmonically bounded candidates

- A "harmonically bounded" candidate is one that can't win under any Classic OT constraint ranking
  - Harmonically bounded candidates also can't win under any grammar that's a probability distribution over Classic OT (partial ranking, stochastic OT)
  - We saw that in Harmonic Grammar, a harmonically bounded candidate can at best tie with other candidates.
  - In non-noisy HG, we can have a straightforward tie of the three candidates in the tableau below.
  - But, in noisy HG, harmonically bounded *cand2* has only an infinitesimal chance of winning:

	CONSTRAINT1 weight: -1 weight+noise: -1+*a*	CONSTRAINT2 weight: -1 weight+noise: -1+*b*	*score*
cand1	**		2(-1+a) = -2+2a
cand2	*	*	(-1+a)+(-1+b) = -2+a+b
cand3		**	2(-1+b) = -2+2b

cand2 wins or ties only if -2+*a*+*b* ≥ -2+2*a*, or *b*≥*a* <u>and</u> -2+*a*+*b* ≥ -2+2*b*, or *a*≥*b*

- So, *cand2* ties for winning when the two noise values *b* and *a* are exactly equal

3 Harmonically bounded candidates in MaxEnt?

- Can a candidate every have a probability of 0 in MaxEnt? Hint: write out the expression for a candidate's probability
- Let's look at the same case again; assume again equally weighted constraints

	CONSTRAINT1 weight: 1	CONSTRAINT2 weight: 1	*probability*
cand1	**		$(e^{-2})/Z) = 0.33$
cand2	*	*	$(e^{-1-1})/Z) = 0.33$
cand3		**	$(e^{-2})/Z) = 0.33$

- How about and even more straightforwardly harmonically-bounded candidate?

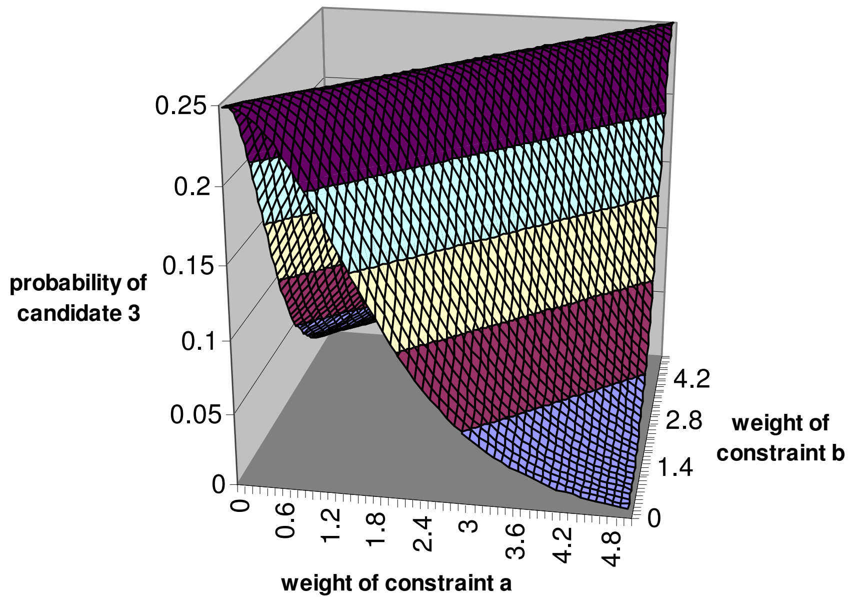
	CONSTRAINT3 weight: 1	CONSTRAINT4 weight: 1	*probability*
cand4	*		$(e^{-1})/Z) = 0.73$
cand5	*	*	$(e^{-1-1})/Z) = 0.27$

4 Multi-site variation

- This brings us back to the questions from Day 1 about multi-site variation:

/mɑɹkətəbɪləti/	*t/V_V[1] weight = *a*	IDENT(continuant) weight = *b*	*probability*
cand1 [mɑɹkətʰəbɪlətʰi]	**		$(e^{-2a})/Z$
cand2 [mɑɹkəɾəbɪləɾi]		**	$(e^{-2b})/Z$
cand3 [mɑɹkətʰəbɪləɾi]	*	*	$(e^{-a-b})/Z$
cand4 [mɑɹkəɾəbɪlətʰi]	*	*	$(e^{-a-b})/Z$

- How does the probability of *cand3* (which is the same as *cand4*), vary as *a* and *b* vary?



- MaxEnt predicts that, if the two constraints' weights are close enough to allow variation between *cand1* and *cand2*, then *cand3* and *cand4* should be strong contenders too.
- So the surprising pattern is when *cand3/cand4* don't occur (see Kaplan's analysis of Warao: there's an additional harmony constraint that rules out *cand3/cand4*).

[1] big simplification

5 Markedness suppression—Kaplan 2012

- Kaplan proposes another quantitative model of variation, designed for multi-site variation.
- If a markedness constraint is designated as suppressible ("⊙"), then each * is subject to being ignored, with some probability *p* that speakers have to learn.
- In this tableau, there are 4 *s under the ⊙ constraint, so there are $2^4 = 16$ possible tableaux. If no marks are suppressed, *cand2* wins:

/mɑɹkətəbɪləti/		⊙*t/V_V	IDENT(continuant)
cand1	[mɑɹkətʰəbɪlətʰi]	**	
cand2	[mɑɹkəɾəbɪləɾi]		**
cand3	[mɑɹkətʰəbɪləɾi]	*	*
cand4	[mɑɹkəɾəbɪlətʰi]	*	*

- Here's a tableau where *cand1* wins.
  - The ◦ indicates that the * has been suppressed
  - In terms of choosing the winner, ◦ is the same as nothing—it's just there to help the reader understand what's happening

/mɑɹkətəbɪləti/		⊙*t/V_V	IDENT(continuant)
cand1	[mɑɹkətʰəbɪlətʰi]	◦◦	
cand2	[mɑɹkəɾəbɪləɾi]		**
cand3	[mɑɹkətʰəbɪləɾi]	*	*
cand4	[mɑɹkəɾəbɪlətʰi]	*	*

probability of this tableau: $p^2(1-p)^2$

- Here's one where *cand3* wins

/mɑɹkətəbɪləti/		⊙*t/V_V	IDENT(continuant)
cand1	[mɑɹkətʰəbɪlətʰi]	◦*	
cand2	[mɑɹkəɾəbɪləɾi]		**
cand3	[mɑɹkətʰəbɪləɾi]	◦	*
cand4	[mɑɹkəɾəbɪlətʰi]	*	*

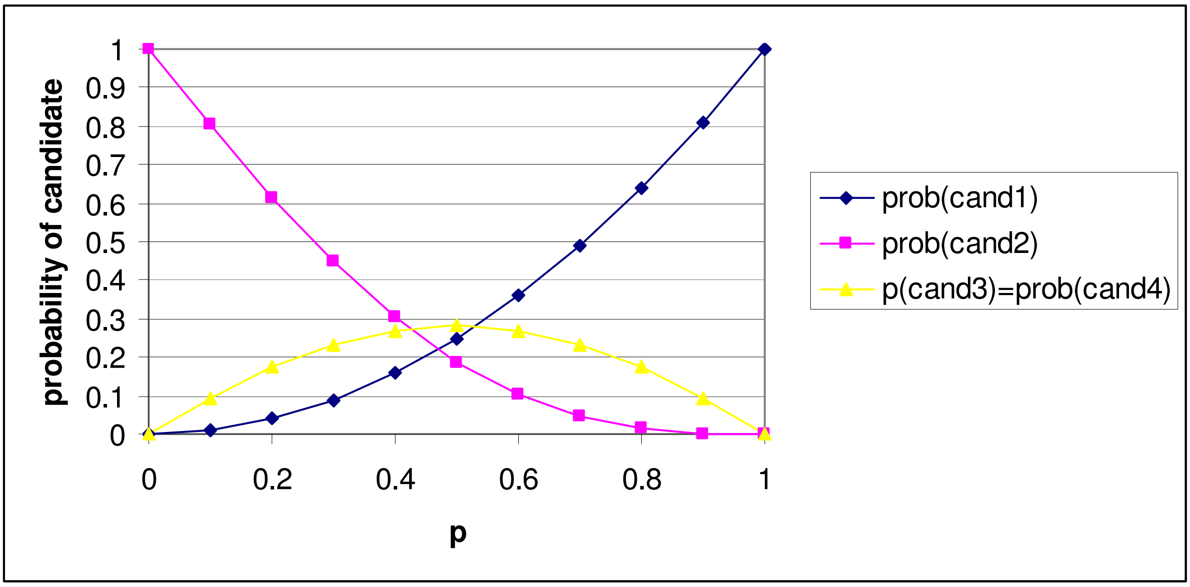
- To find out how probable each candidate is, we need to add up the probabilities of the tableaux that will choose them.
  - Here's a table of each possible suppression pattern for ⊙*t/V_V

cand1	**	◦*	*◦	**	**	◦◦	◦*	◦*	*◦	*◦	**	◦◦	◦◦	◦*	*◦	◦◦
cand2																
cand3	*	*	*	◦	*	*	◦	*	◦	*	◦	◦	*	◦	◦	◦
cand4	*	*	*	*	◦	*	*	◦	*	◦	◦	*	◦	◦	◦	◦
winner	2	2	2	3	4	1	3	4	3	4	3/4 tie	1	1	3/4 tie	3/4 tie	1
prob. of tableau	$(1-p)^4$	$p(1-p)^3$				$p^2(1-p)^2$						$p^3(1-p)$				p^4
e.g., if *p*=0.2	0.410	0.102	0.102	0.102	0.102	0.026	0.026	0.026	0.026	0.026	0.026	0.006	0.006	0.006	0.006	0.002

- So, for *p*=0.2, the probabilities of the candidates are as follows (assume equal split when tied):

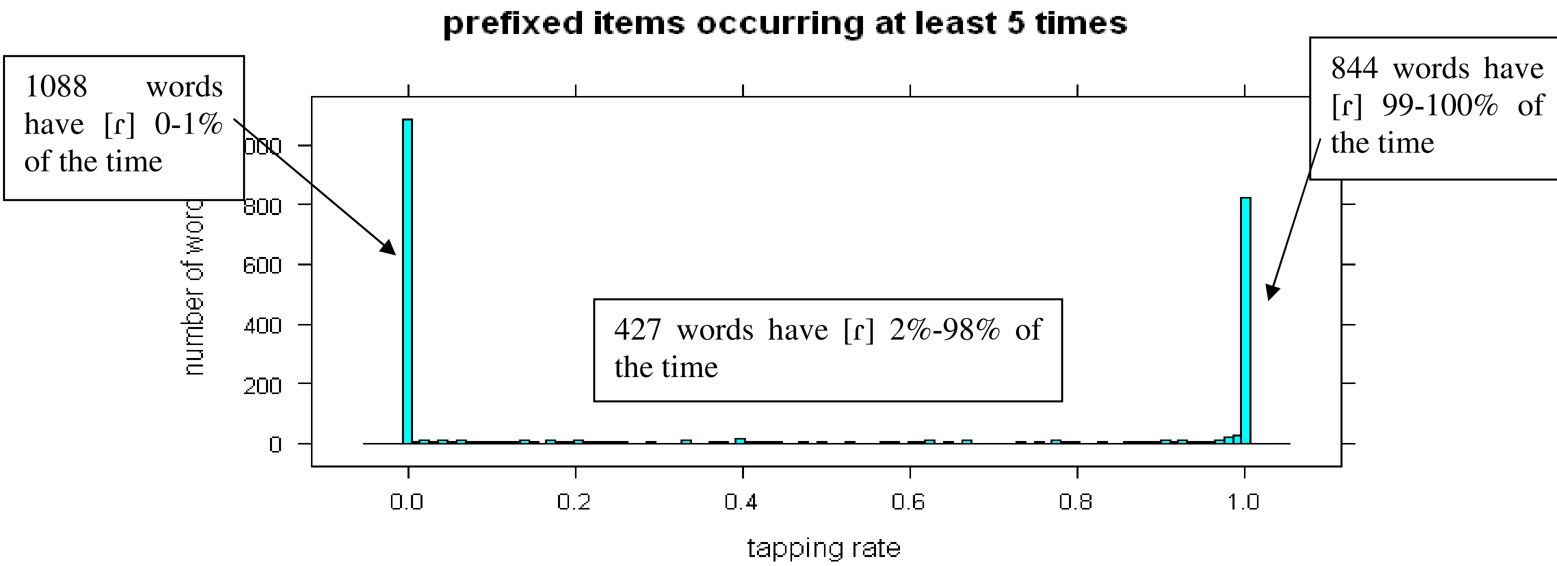
	probability	
cand1	0.026+0.006+0.006+0.002	= 0.04
cand2	0.410+0.102+0.102	= 0.61
cand3	0.102+0.026+0.026+(0.026/2)+(0.006/2)+(0.006/2)	= 0.17
cand4	0.102+0.026+0.026+(0.026/2)+(0.006/2)+(0.006/2)	= 0.17

- We can plot how the probabilities of the candidates change as *p* changes:



6 Lexical variation

- We've focused on free variation—or pretended that lexical variation was really free variation—because it's easier.
- Recall the Tagalog case of d → ɾ / V__V.
  - If we look at each prefixed word, like /ma+**d**umi/, determine its rate of undergoing the change, and then count up how many words have each rate, we see a strong skewing towards 0% and 100%:



==> Most words have a fixed behavior, though some do vary

7 Modeling lexical variation: indexed constraints

- Probably the best-developed theory of lexical variation is **constraint indexing** (Pater 2009, Becker 2009, Mahanta 2009[2])

[2] There are earlier references from the same authors, but I chose works that seemed to represent the most current versions of the authors' approaches.

- The basic idea, as applied to our Tagalog case:
  *VdV$_{\text{type A words}}$ >> IDENT(cont) >> *VdV$_{\text{type B words}}$
  - Let's draw tableaux for /ma/+/dunoŋ/$_A$ and /ma/+/daʔig/$_B$

- Richer example, from Becker 2009: Turkish (Altaic language from Turkey with 50 million speakers, Ethnologue 2005)
- Three kinds of word-final obstruents in Turkish (p. 19)

always voiceless

ɑtʃ	'hunger'	ɑtʃ-ɨ	'hunger-*possessive*'
ɑnɑtʃ	'female cub'	ɑnɑtʃ-ɨ	'female cub-*possessive*'

always voiced (rarer—examples from Kaisse 1990*)*

ofsɑjd	'offside'	ofsɑjd-ɨ	'offside-*possessive*'
serhɑd	'Serhad (name)'	serhɑd-ɨ	'Serhad's'

alternating

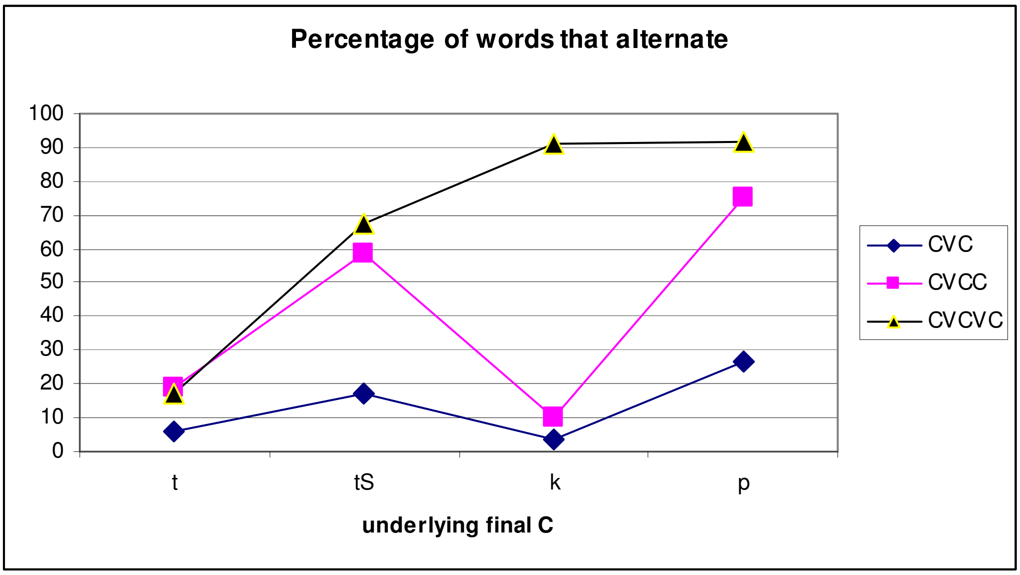
tɑtʃ	'crown'	tɑdʒ-ɨ	'crown-*possessive*'
ɑmɑtʃ	'target'	ɑmɑdʒ-ɨ	'target-*possessive*'

  - Becker takes the unaffixed form as underlying (this is different from the classic devoicing analysis)
  - The grammar then needs to let some words undergo intervocalic voicing.

  - Let's develop an indexed-constraint analysis of the Turkish data so far.

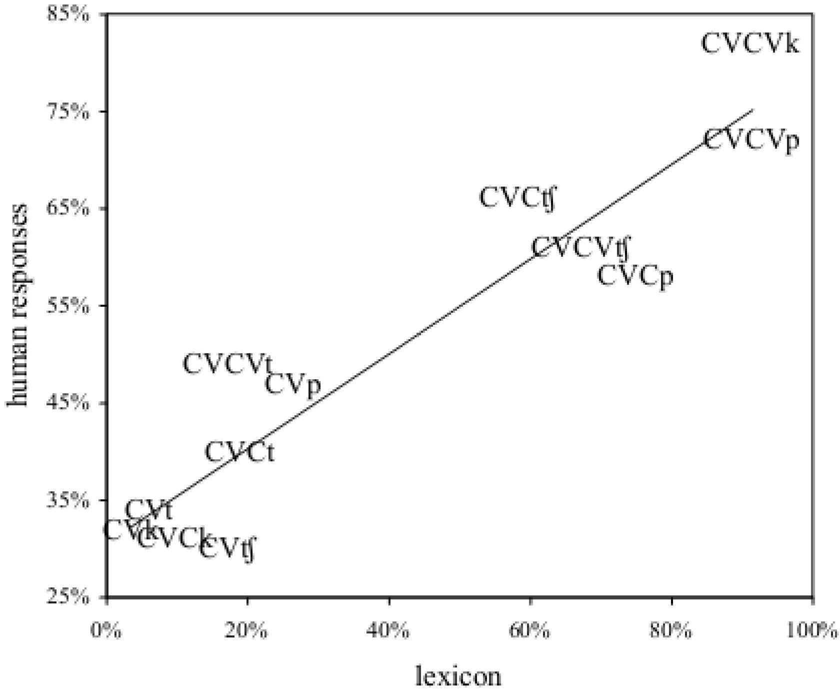
8 Patterned exceptions

- This isn't enough, because distribution of always-voiceless vs. alternating isn't random.



Based on Becker's p. 25

- However this came about historically, Turkish speakers seem to have learned the pattern.
  - In a wug-test (Berko 1958), speakers followed the pattern closely, though overall closer to 50-50:



(Becker p. 37)

==> We want to get this information into the grammar

9 Grammar for Turkish

- Becker proposes that Turkish learners have access to these constraints:
  - IDENT(voice)
  - IDENT(voice)$_{\sigma 1}$: the [voice] value of an output segment in a word's first syllable must be the same as the [voice] value of its input correspondent
  - *VtV, *VtʃV, *VpV, *VkV
  - *RtV, *RtʃV, *RpV, *RkV : don't have [p] (etc.) preceded by a sonorant C and followed by a vowel
- The learner encounters an inconsistency...
  - /ɑnɑtʃ-ɨ/ → [ɑnɑtʃ-ɨ] means IDENT(voice) >> *VtʃV
  - /ɑmɑtʃ-ɨ/ → [ɑmɑdʒ-ɨ] means *VtʃV >> IDENT(voice)
- So, the learner **clones** the IDENT constraint and re-does the ranking (see Becker's ch. 4 for how to choose which constraint to clone).
  - IDENT(voice)$_{\text{<*VtʃV,ɑnɑtʃ>}}$: Don't change voicing values in the lexical item /anatʃ/; conflicting constraint is *VtʃV
  - IDENT(voice)$_{\text{<*VtʃV,ɑmɑtʃ>}}$

- Let's make tableaux for /ɑnɑtʃ-ɨ/, /ɑmɑtʃ-ɨ/, with the two IDENT clones and *VtʃV

- Eventually, the learner ends up with constraints like...
  IDENT(voice)$_{\text{\{<*VtʃV,ɑnɑtʃ>, <*VtV,sepet>,...\}}}$ (49 <*VtʃV, *X*> items)
  IDENT(voice)$_{\text{\{<*VtʃV,ɑmɑtʃ>, <*VtV,kɑnɑtt>, ...\}}}$ (101 <*VtʃV, *X*> items)
- When it's time to take the wug test, experimental participant must choose which IDENT constraint to assign the new word /hevetʃ/, with the conflicting constraint being *VtʃV.
- The more existing <*VtʃV,X> items belong to the constraint, the more likely the new word is to be assigned to it.

- As far as I know there's no software available for implementing this learner yet.

10 A different model (see Zuraw 2010 for details and some discussion of learning[3])

- Suppose that Turkish speakers just have lexical entries all the affixed words they know: /ɑnɑtʃ-ɨ/, /ɑmɑdʒ-ɨ/

- Known words surface faithfully—I'm illustrating this with Stochastic OT ranking values (how faithfulness works here is a bit of a simplification—see the paper for more details)

input: /ɑnɑtʃ-ɨ/, O-O corr to /ɑnɑtʃ/	**IDENT-IO(voice) R.V.: 110**	*VpV 98	IDENT-OO(voice) 97.5	*VtʃV 97
☞ *a* [ɑnɑtʃ-ɨ]				*
b [ɑnɑdʒ-ɨ]	*		*	

- But new words are subject only to lower-ranking constraints, because there's nothing to be faithful to:

no suffixed form exists in lexicon O-O corr to /hevetʃ/	IDENT-IO(voice) R.V.: 110	***VpV 98**	**IDENT-OO(voice) 97.5**	***VtʃV 97**
62% *c* [hevetʃ-ɨ]				*
38% *d* [hevedʒ-ɨ]			*	

no suffixed form exists in lexicon O-O corr to /hevep/	IDENT-IO(voice) R.V.: 110	***VpV 98**	**IDENT-OO(voice) 97.5**	***VtʃV 97**
27% *e* [hevep-ɨ]		*		
73% *f* [heveb-ɨ]			*	

- o In this model, how can we rule out pairs like hypothetical /sat/ 'frisbee' /fim-ɨ/ 'frisbee-*poss*'?
- o In this model, how can we ensure that the various suffixed forms of the same stem all have the same voicing behavior?

11 Variation and grammar architecture

- English *t/d* deletion: *belt* [bɛlt]~[bɛl], *felt* [fɛlt]~[fɛl], *clapped* [klæpt] ~[klæp], etc.
- As you read about in the Coetzee article, this phenomenon has a long history of sociolinguistic study, starting with Labov
- Has been described for many dialects, which have different overall rates of deletion but similar sensitivity to conditioning factors:
- Phonological conditioning
  - __#V vs. __#C vs. __ *pause*
  - preceding C (especially, how similar to *t*/*d*)
  - target *t* vs. *d*

[3] MaxEnt didn't work too well here for learning both invariance of listed items and overall phonological trends; GLA/Stochastic OT worked better. But, I later found that for learning the differences between morphemes, GLA did poorly and MaxEnt was better (unpublished); I should try Magri's version of GLA.

- Morphological conditioning
  - monomorphemes: *belt, weld, sand, tend, mist*
  - "semi-weak past"—vowel quality changes but suffix is also added: *kept, wept, slept, felt, meant, told, left...*
  - regular past: *slapped, wrapped, healed, missed*
- Guy (1991a,b)—who I think was the first to notice the difference between semi-weak and regular past, though I'm not sure—relates this to his previous proposal of Lexical Phonology (Kiparsky 1982, Mohanan 1986 and others) plus variation.

Let's see how this works...

- Lexical Phonology, ignoring *t/d* deletion

		derivation of *mist*	derivation of *slept*	derivation of *missed*
	lexical entry	/mɪst/	/slip/	/mɪs/
lexical component	LEVEL I: irregular inflection	--	slɛpt	--
	LEVEL II: regular inflection	--	--	mɪst
	postlexical	--	--	--

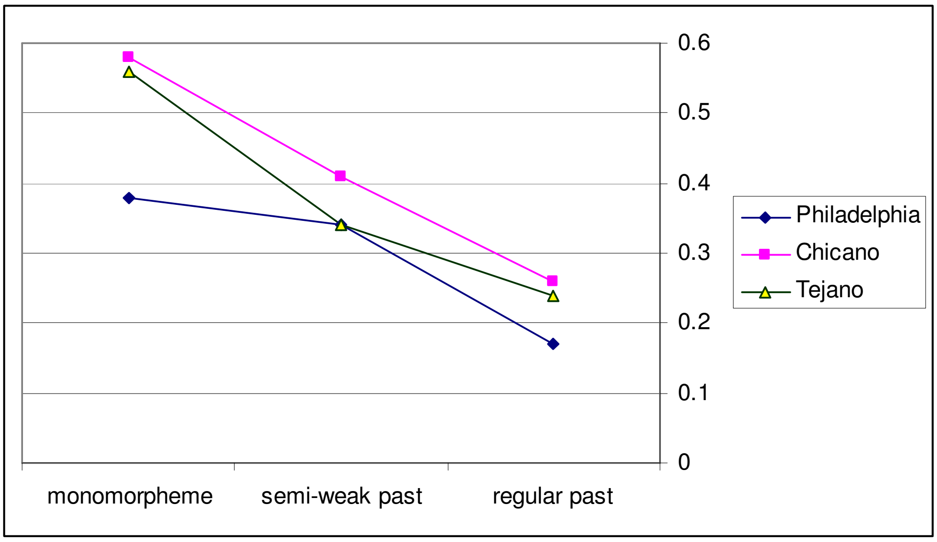
  - Only for adults: children, adolescents, and younger adults may treat words like *slept* as monomorphemic

- Suppose that *t/d* deletion is a variable rule, with a probability *p* of applying:

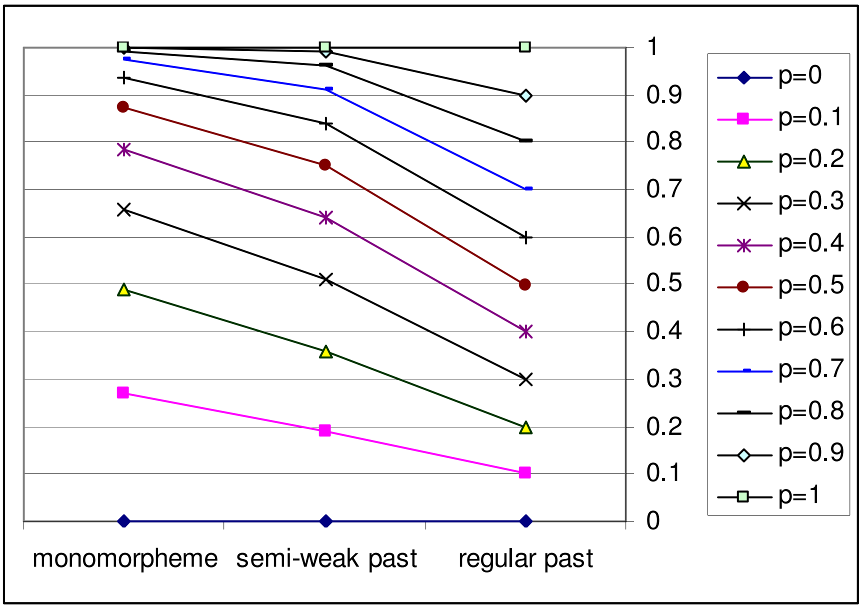
	derivation of *mist*	derivation of *slept*	derivation of *missed*
lexical entry	/mɪst/	/slip/	/mis/
variable *t/d* deletion	chance of deletion	--	--
LEVEL I: irregular inflection	--	slɛpt	--
variable *t/d* deletion	chance of deletion	chance of deletion	--
LEVEL II: regular inflection	--	--	mɪst
variable *t/d* deletion	chance of deletion	chance of deletion	chance of deletion
postlexical	--	--	--
probability of deletion	$1-(1-p)^3$	$1-(1-p)^2$	$1-(1-p)$
e.g., if p=*0.2*	0.49	0.36	0.20

Let's derive this part on the board

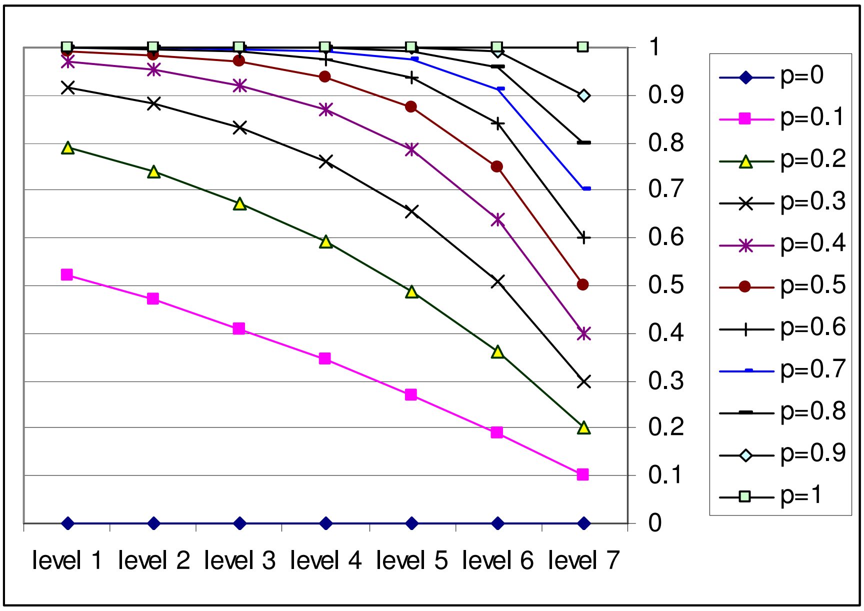
- Abstracting away from the effect of the phonological context, here are the data for three dialects of English, as you read in Coetzee (data from Guy 1991, Bayley 1995, Santa Ana 1992, see also Santa Ana 2008):



- Here's what the model predicts for different values of *p*:



- What would it look like if there were even more levels?



- See McPherson & Hayes 2012 for a case with more than 3 “levels”, where the data pattern doesn’t seem to follow this exponential pattern.

12 Which model of variation in OT?

- Kiparsky 1993 implements each level as a partially ordered OT grammar, but we could substitute a different variable constraint model with the same results
  - What matters is just the probability of (in this case) deletion that the grammar predicts.

13 Gradience vs. variability: Myers 1995

- Myers suggests that we pay more attention to the difference between...
  - gradience: /t/ could have 100 msec duration, 80 msec, 20 msec, 0 msec...
  - variability: /t/-deletion rule applies (presumably deleting the /t/ completely—0 msec) or doesn’t apply
- See his paper for some very interesting discussion (and data) of predicted patterns

14 Course summary—I’ll keep it very brief!

- We’ve seen different aspects of variability that could be problematic
  - free vs. lexical vs. mixed
  - multi-site: do the sites vary independently or are they related?
- We spent some time on regression models
  - helpful for exploring data
  - helpful for reading experiment literature, literature from psychology, etc.
  - well-developed software and math for significance testing, smoothing, etc.
- This wasn’t a real statistics course, though.
  - If you want to do statistical analysis on a serious project, you could use what you learned here to make a preliminary model,
  - then hire a statistics grad student to give you advice on things like whether your data meet the assumptions of the regression, whether you need to center your variables, what kinds of smoothing/prior you should use (e.g., the bayesGLM() function in R), model comparison, etc.

- We saw various constraint models of variation
  - partial ordering
  - Stochastic OT
  - noisy Harmonic Grammar
  - maximum entropy constraint grammar ≈ logistic regression
- We talked about the general problem of overfitting vs. underfitting; model comparison
  - well-developed topic in statistics: “smoothing”, “regularization”, “priors”
  - in constraint models, well developed only in MaxEnt (Gaussian prior)
- Finally, we took a look at variability beyond the tableau level, such as how phonology should interact with morphology
- Some open questions
  - What are the empirical facts? Which of these models actually works best?
  - We would like to have better data for...
    - phonology-phonology interaction
    - phonology-morphology interaction
    - multi-site variation cases
    - gradience (in Myers’s sense)

Next time: Your presentations! We’ll see what you’ve been working on or are considering working on.

References

Bayley, Robert. 1995. Consonant cluster reduction in Tejano English. *Language Variation and Change* 6:303-326.

Becker, Michael. 2009. Phonological trends in the lexicon: the role of constraints. University of Massachusetts Amherst Ph.D. dissertation.

Berko, Jean. 1958. The child’s learning of English morphology. *Word* 14. 150-177.

Coetzee, Andries W. 2009. An integrated grammatical/non-grammatical model of phonological variation. In Young-Se Kang, Jong-Yurl Yoon, Hyunkung Yoo, Sze-Wing Tang, Yong-Soon Kang, Youngjun Jang, Chul Kim, Kyoung-Ae Kim, & Hye-Kyung Kang (eds.), *Current Issues in Linguistic Interfaces*, vol. 2, 267–294. Seoul: Hankookmunhwasa.

Ethnologue. 2005. *Ethnologue: Languages of the World*, Fifteenth edition. (Ed.) Raymond G. Gordon.

Guy, Gregory R. 1991. Explanation in Variable Phonology: An Exponential Model of Morphological Constraints. *Language Variation and Change* 3: 1–22.

Guy, Gregory R. 1991. Contextual conditioning in variable lexical phonology. *Language Variation and Change* 3: 223-239.

Kaplan, Aaron F. 2011. Variation through Markedness Suppression. *Phonology* 28: 331-370.

Kiparsky, Paul. 1982. Lexical morphology and phonology. In I. S. Yang (ed.), *Linguistics in the morning calm*. Seoul: Hanshin. 3-91.

Kiparsky, Paul. 1993. An OT perspective on phonological variation. Handout from Rutgers Optimality Workshop 1993, also presented at NWAVE 1994, Stanford University.

Magri, Giorgio. to appear. HG has no computational advantages over OT: towards a new toolkit for computational OT. *Linguistic Inquiry.*

Mahanta, Shakuntala. 2009. Morpheme-specific exceptional processes and emergent unmarkedness in vowel harmony.. In Rajendra Singh (ed.), *Annual review of South Asian languages and linguistics: 2009*. Walter de Gruyter.

McPherson, Laura & Bruce Hayes. 2012. Relating application frequency to morphological structure: the case of Tommo So vowel harmony. Talk presented at Manchester Phonology Meeting.

Mohanan, K. P. (1986). *The theory of lexical phonology*. Dordrecht: Reidel.

Myers, James. 1995. The categorical and gradient phonology of variable t-deletion in English. Paper presented at the International Workshop on Language Variation and Linguistic Theory, University of Nijmegen, Netherlands.

Pater, Joe. 2009. Morpheme-specific phonology: constraint indexation and inconsistency resolution. In Steve Parker (ed.), *Phonological argumentation: essays on evidence and motivation*. (Advances in Optimality Theory). Equinox.

Santa Ana, Otto. 1991. Phonetic Simplification Processes in the English of the Barrio: A Cross-Generational Sociolinguistic Study of the Chicanos of Los Angeles. Ph.D. dissertation, University of Pennsylvania.

Santa Ana, Otto. 2008. Chicano English evidence for the exponential hypothesis: A variable rule pervades lexical phonology. *Language Variation and Change* 4: 275-288.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *Natural Language and Linguistic Theory* 28: 417-472