Learning hidden generalizations

24.964—Fall 2004
Modeling phonological learning

Class 10 (18 Nov, 2004)

Agenda for today

“Hidden” generalizations, seen from three perspectives

- Continued discussion of Boersma (2004)
- Everett and Berent (1997)
- Zuraw (2000)
- (If time:) Some notes on evaluating models (held over from Week 3)

Boersma (2004)

Boersma (2004)

Reminder of part 1: reply to Keller and Asudeh

- Points out that the K&A examples are mostly irrelevant, because they concern comparisons between outputs *for different inputs*
  - ○ GLA does not aim to model why some words are more frequent than others
  - ○ It is a model of competing variants for a single input

- Acknowledges an interesting puzzle: why are things which are unattested sometimes somewhat acceptable?
  - ○ Proposes that losing candidates can be compared with each other
  - ○ Similar to proposal by Coetzee (2004)

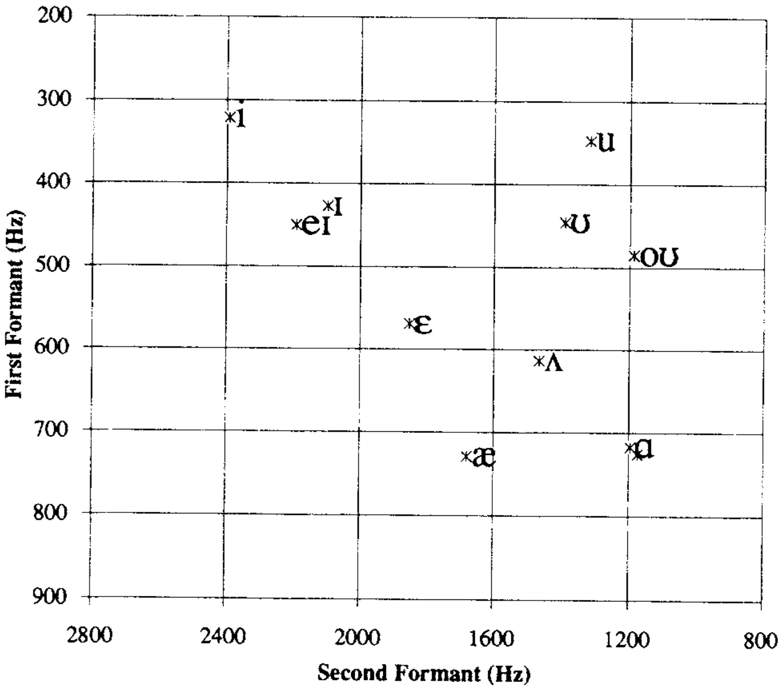
Boersma (2004)

Background on prototypes: Johnson, Flemming and Wright (1993)

- Speakers tend to prefer (rate as better) vowels that are more extreme in the vowel space than what they would actually produce

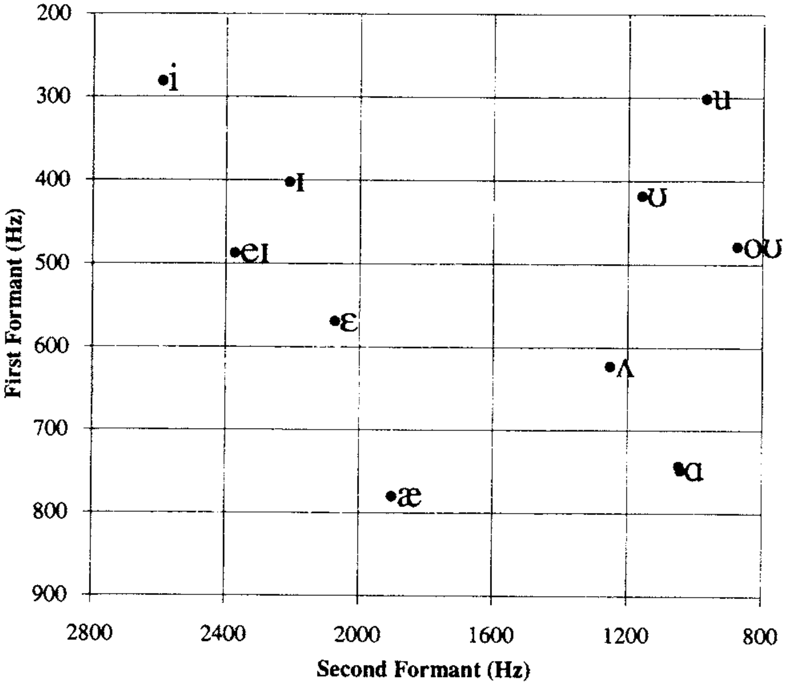
Johnson, Flemming and Wright (1993)

Speakers’ productions (averaged over 8 male speakers)



Johnson, Flemming and Wright (1993)

The vowels these same speakers pick as “best”



Johnson, Flemming and Wright (1993)

Proposed analysis:

- /UR/ → [SR] → Ideal, hyperarticulated realization → Actual implementation

Boersma (2004)

Clunkiness of the quantization

- "an F1 of 340 Hz is not /a/", "an F1 of 330 Hz is not /a/", etc.

Seems like we would really want the constraints to be continuous functions, mapping frequency to probability

- Might Bayesian inversion help here?
  - $\mathrm{P}(/\mathrm{i}/ \mid \mathrm{F1} = n\,\mathrm{Hz}) = \frac{\mathrm{P}(\,\mathrm{F1} = n\,\mathrm{Hz} \mid /\mathrm{i}/) \times \mathrm{P}(\,/\mathrm{i}/)}{\mathrm{P}(\,\mathrm{F1} = n\,\mathrm{Hz})}$

Everett and Berent (1997)

Everett and Berent (1997)

p. 12
“If SSM outputs correspond to SSM inputs, but SMM outputs correspond to SM inputs (as they would do, based on markedness and learnability) then the derivations of the relevant forms entail different numbers of violations of the OCP. SSM-type forms violate the OCP at both the deep and the intermediate level, leading to an accumulation of at least two violations of the OCP. Forms such as maSSiMim, where the geminates are also adjacent at the surface level, manifest a third violation of the OCP. Conversely, MSS-type forms [xxx or SMM] do not violate the OCP at the deep level, because their input is MS [xxx SM]. However, in preparation for plan conflation, their rightmost radical reduplicates, thereby violating the OCP at the intermediate level. Thus, MSS-type forms accumulate only one violation of the OCP.”

Everett and Berent (1997)

A derivational account:

		Deep	Intermediate	Surface
SSM	sisem	*	*	
SSM	massimim	*	*	*
SMM	simem	*	*	
SM	simem		*	
PSM	pisem			

Everett and Berent (1997)

The comparative account: (p. 14)

OUTPUT		OCP	*Initial ID	*Final ID
word 1	sisem		*!	
word 2	massimim	*!	*	
word 3	simem			*
word 4	pisem			

Everett and Berent (1997)

The comparative account: (p. 14, fixed up)

OUTPUT		OCP	*Initial ID	*ID
word 1	sisem		*	*
word 2	massimim	*	*	*
word 3	simem			*
word 4	pisem			

Everett and Berent (1997)

What does this theory mean for the concept of ‘ungrammaticality’?

- massimim is generated by the grammar, but it’s unacceptable to speakers

…and what does this mean for Richness of the Base?

Everett and Berent (1997)

The learning issue involved:

- This all rests crucially on the fact that speakers can represent [simem] as /SM/ (rather than /SMM/
- If you know the relative ranking of the OCP, then it's not a problem; /SMM/ and /SM/ both yield [simem] (somehow), but /SM/→[simem] is more harmonic (no OCP violation)
- But why don't learning data like [simem] inspire the learner to demote the OCP, so that it doesn't play a role in the adult grammar?

Putting these together

- Boersma (2004), Coetzee (2004): all candidates are ranked harmonically, allowing us to adjudicate between relative well-formedness of losers
  - ○ Boersma: attempt to translate to numerical predictions
- Everett and Berent (1997): all words are ranked harmonically
  - ○ Allows us to compare relative well-formedness of outcomes for different lexical items

Zuraw (2000)

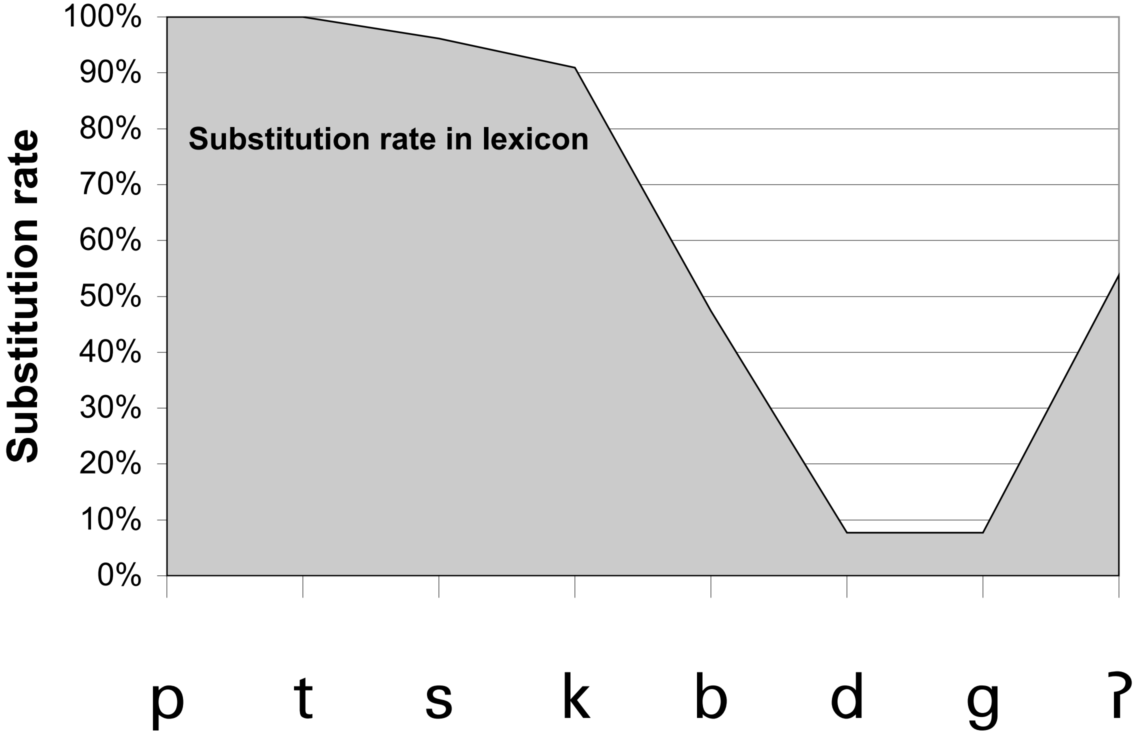
Reminder of the data: Tagalog nasal substitution

	pighatíʔ	‘grief’
	pa-mi**m**ighatíʔ	‘being in grief’
vs.		
	poʔók	‘district’
	pam-**p**oʔók	‘local’

(See Zuraw, chapter 2, ex. 7 for more examples)

Zuraw (2000)

Exceptions are not distributed evenly for all consonants



- These counts are for /maŋ+ R_{CV}/; see Zuraw (2000) for other prefixes, and pooled across prefixes

Zuraw (2000)

The role of voicing:

OBSERVED	unsub	sub	total
vcls	46	578	624
vcd	217	142	359
total	263	720	983

EXPECTED	unsub	sub	total
vcls	166.95	457.05	624
vcd	96.05	262.95	359
total	263	720	983

$$\chi^2 = 327.572, p < .0001$$

Zuraw (2000)

The role of place:

- Labial > coronal (marginally sig. for vcls, very sig. for vcd)
- Coronal > velar (n.s. for vcls, very sig. for vcd)

Zuraw (2000)

An experiment to test the productivity of nasal substitution

Pagbubugnat	ang	trabaho	niya.	Siya	ay ____
bugnat-ing	TOP	job	POSS	he/she	

His/her job is to *bugnat*. He/she is a ____

“*mangbubugnat* / *mambubugnat* / *mamumugnat* / ???”

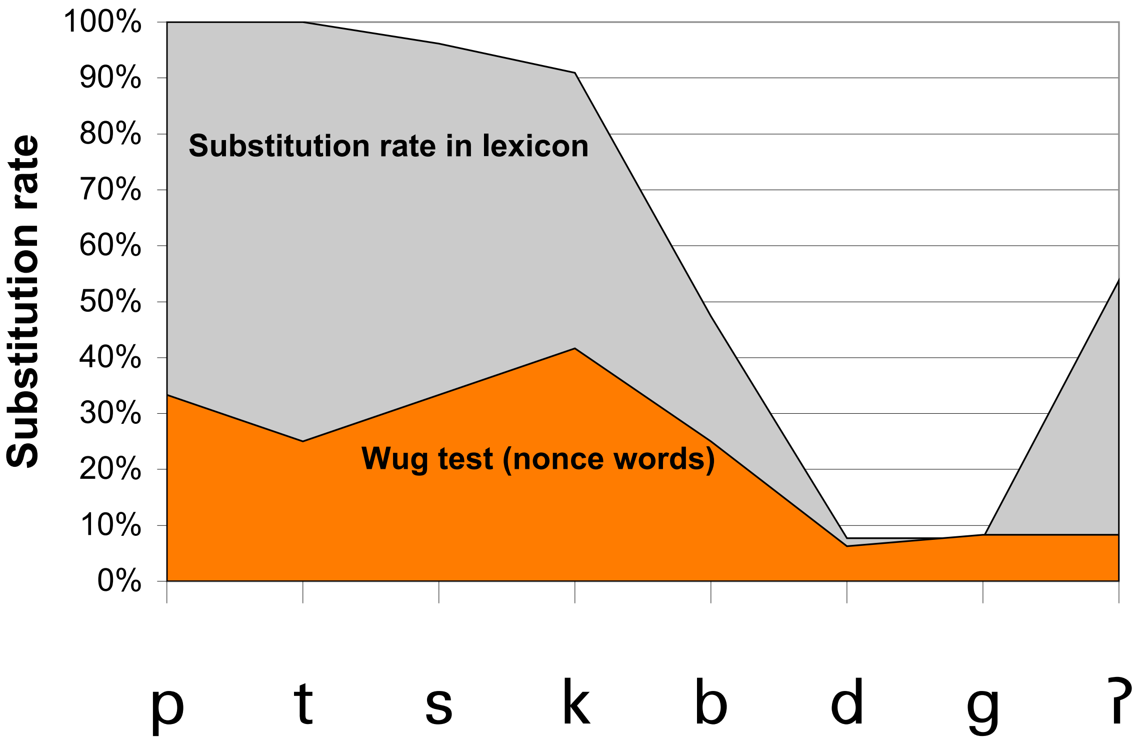
Zuraw (2000)

Two groups:

- Group A: Instructions claimed that these were real (but rare) words
- Group B: Instructions said they were made up, no right or wrong answers

Zuraw (2000)

Reminder of results: productivity across consonants generally mirrors rate of substitution for existing words



Zuraw (2000)

Results: overall, substitution rates were low for both groups

Group B: makes sense preserve integrity of root, to aid recoverability)

Group A: Zuraw suggests maybe following pattern specifically of rare words (is this so different from treating as a nonce words? need to promote recoverability in either case)

Zuraw (2000)

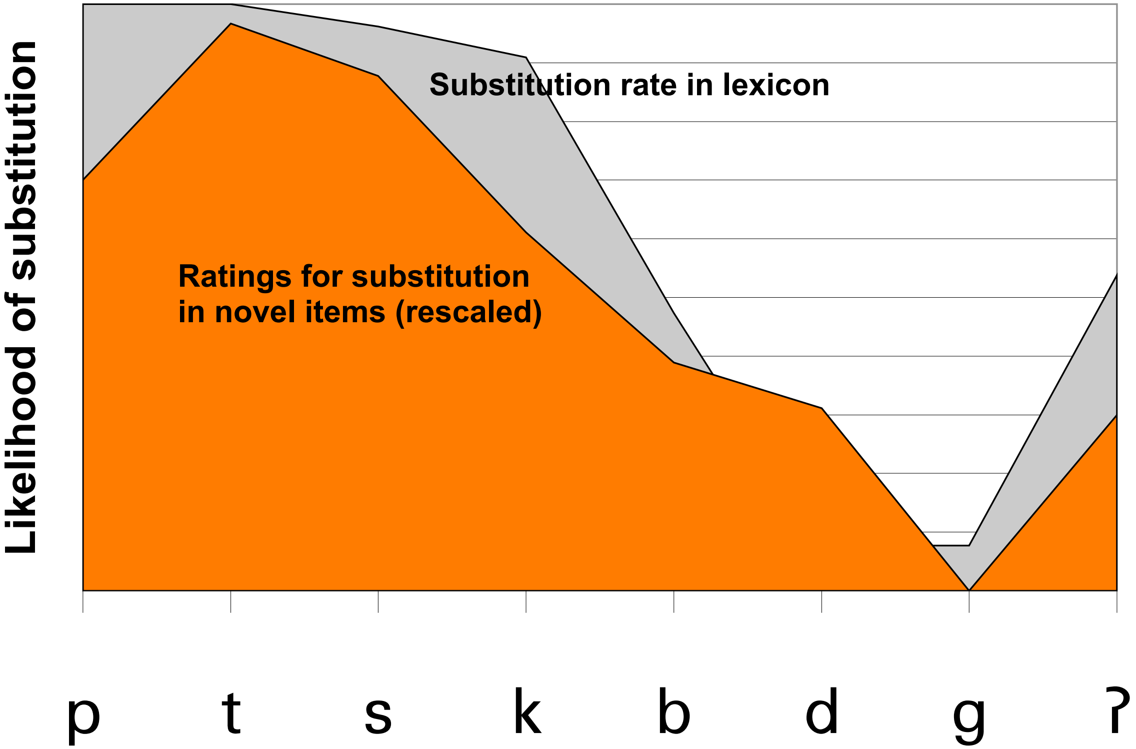
Task 2, designed to get more detailed data

- Ratings task, on substituted and unsubstituted items
- Participants rated from 1 (bad) to 10 (good)
- All participants rated both substituted and unsubstituted, so it's possible to calculate the difference = degree of preference (or dispreference) for substiitution

Zuraw (2000)

Results: ratings follow lexical trends

- Ratings estimated and rescaled from Zuraw's figures, for overlaid comparison



Zuraw (2000)

So how do we analyze this?

- Nasal substitution = coalescence
  - /m$_{1}$ a$_{2}$ ŋ$_{3}$ + b$_{4}$ i$_{5}$ g$_{6}$ a$_{7}$ j$_{8}$ / → [m$_{1}$ a$_{2}$ m$_{3,4}$ i$_{5}$ g$_{6}$ a$_{7}$ j$_{8}$]
- Coalescence across a morpheme boundary violates MorphOrder
  - If morpheme μ_{1} precedes μ_{2} in the input, all segs of μ_{1} must precede μ_{2} in the output
- Coalescence within a listed entry violates EntryLinearity
  - If seg X precedes seg Y in the lexical entry, then the surface correspondent of X must precede the surface correspondent of Y in the ouput
- OO constraints: penalize substitution (since it creates alternations)

Zuraw (2000)

Many possible parses/inputs to consider:

(full list on p. 49)

‘to *bigaj*’	ID [place]	ID [son]	DEP	MAX	MORPH ORDER	ENTRY LIN
/maŋ$_3$+/b$_4$igaj/ → [mam$_{3,4}$igaj]	*	*			*	
/maŋ$_3$+/b$_4$igaj/ → [mam$_3$igaj]	*			*		
/maŋ$_3$+/b$_4$igaj/ → [mam$_4$igaj]		*		*		
/maŋ$_3$+/b$_4$igaj/ → [mam$_3$b$_4$igaj]	*					
/mam$_3$i$_4$gaj/ → [mam$_3$i$_4$gaj]						
/mam$_3$i$_4$gaj/ → [mam$_3$bi$_4$gaj]			*			
/mam$_3$i$_4$gaj/ → [mam$_3$b$_3$i$_4$gaj]		*				*

(etc.)

Zuraw (2000)

How do we know whether we're dealing with a complex input (/maŋ+bigaj/) or a simple one (/mamigaj/ or /mambigaj/)?

- Both are possibilities
- But there is a preference to use single lexical entries
  - USELISTED: the input of evaluation must be a single lexical entry
- Similar in spirit to the blocking principle (Aronoff 1976)

Zuraw (2000)

Illustration of UseListed, p. 51

	Meaning	UseListed
a. /mamigaj/ → …		
b. /maŋ/+/bigaj/ → …		*
c. /ʔipamigaj/ → …	*	
d. /mag/+/bigaj/ → …	*	*

- An implementation of morphological blocking; ensures that specific lexical entry is used, if one exists with the right meaning

Zuraw (2000)

One other refinement: strength of listing

- "Listedness" may vary from 100% (completely listed) to 0% (not listed at all)
- Can interact with OO (paradigm uniformity) constraints

Zuraw (2000)

Interaction of UseListed and Paradigm Uniformity, p. 53

	Entry Lin	Use 40% Listed	PU	Use 30% Listed
1a. /manala/$_{\text{30\% listed}}$ → [manala]		*	*!	
1b. /manala/$_{\text{30\% listed}}$ → [mantala]	*!	*		
1c. /maŋ/+/tala/ → [manala]		*	*!	*
1d. ☞ /maŋ/+/tala/ → [mantala]		*		*
2a. ☞ /manili/$_{\text{40\% listed}}$ → [manili]			*	
2b. /manili/$_{\text{40\% listed}}$ → [mansili]	*!			
2c. /maŋ/+/sili/ → [manili]		*!	*	*
2d. /maŋ/+/sili/ → [masnili]		*!		*

Zuraw (2000)

Segment-by-segment differences

- The voicing effect
  - *NC̥
- The place effect
  - *[ŋ≫ *[n ≫ *[m

Zuraw (2000)

The meat of the analysis, part 1

- Tagalog allows both substituting and non-substituting words—thus, faithfulness must be ranked high enough to allow both to surface, if listed

(see p. 62 for relevant tableaux)

Zuraw (2000)

The meat of the analysis, part 2

- However, when there is no listed form (nonce/rare word), lower ranked constraints get a say

(see p. 63 for schematic example)

Zuraw (2000)

How do you learn, when they are so many possible parses for the input?

- The grammar of Tagalog differs depending on whether the word is listed or not

- In order to learn this ranking, you need to know whether words in the input data are listed or not

- But how on earth do you know whether another speaker was using a listed lexical entry?

Zuraw (2000)

A standard assumption: learner assumes everything's listed

- At first, faithfulness constraints are ranked together with everything else, so GLA starts pushing around all the relevant constraints
- The relevant markedness constraints try to arrange themselves, pushing in conflicting ways in response to conflicting data
- At the same time, faithfulness constraints are always right (since we're operating under the IN=OUT assumption)
- So soon, faithfulness constraints climb to the top, no more errors are made, and learning ceases

Zuraw (2000)

A big problem: how do we test this grammar to see what it predicts?

- Probability of an output depends both on the probability of the rankings that would derive it, and also on the probability of the inputs

see Zuraw (2000), p. 83 for illustration

Zuraw (2000)

One last issue: parsing/listening

- For the learner, it is reasonable to (at least start by) assuming IN=OUT
- For the adult speaker, the issue is more determining whether someone is saying a form of a word that you already know

Zuraw (2000)

Suppose you hear [mamumuntol]
Three possibilities:

- P/maŋ+R_{CV}/+/puntol/ | [mamumuntol])

  = "the prob. that the speaker constructed form compositionally, given that the output was [mamumuntol]"

- P(/mamumuntol/ | [mamumuntol])

  = "the prob. that the speaker used a listed entry /mamumuntol/, given that the output was [mamumuntol]"

- P(/mampupuntol/ | [mamumuntol])

  = "the prob. that the speaker used a listed entry /mampupuntol/, given that the output was [mamumuntol]"

Zuraw (2000)

Suppose you hear [mamumuntol]

Bayesian inversion:

- P/maŋ+R_{CV}/+/puntol/ | [mamumuntol])

$$= \frac{\text{P([mamumuntol] |/maŋ+R}_{\text{CV}}\text{/+/puntol/)} \times \text{P(/mamumuntol/)}}{\text{P([mamumuntol])}}$$

- P(/mamumuntol/ | [mamumuntol])

$$= \frac{\text{P([mamumuntol] | /mamumuntol/)} \times \text{P(/mamumuntol/))}}{\text{P([mamumuntol])}}$$

- P(/mampupuntol/ | [mamumuntol])

$$= \frac{\text{P([mamumuntol] | /mampupuntol/)} \times \text{P(/mampupuntol/)}}{\text{P([mamumuntol])}}$$

(denominators equal across all competing possibilities)

Zuraw (2000)

But how do you know the relative probabilities of different inputs?

- Could just assume based on your own lexicon
- A more sensible strategy: estimate based on a function of the probability of the lexical item, and what you know about the overall productivity of the pattern that would produce it
- Logistic function:

$$\mathrm{P}(/\mathrm{ma\eta}/+/\mathrm{R}_{CV}/+/\mathrm{puntol}/ = \frac{1}{(1+e^{-3+6\times Listedness(wholewd)})\times(1+e^{3-6\times Productivity(mang+Rcv)})}$$

Zuraw (2000)

Estimating productiving of /maŋ+R$_C V$/

- Coverage among known stems (most things have a mang+RED form)
- Lack of correlation between stem forms and mang+RED forms (?)
- Phonological or semantic idiosyncrasy (see also Hay on this point)

Zuraw (2000)

End result of all this

p. 95:

P/maŋ+R_{CV}/+/puntol/ | [mamumuntol]) = .062
P(/mamumuntol/ | [mamumuntol]) = .939
P(/mampupuntol/ | [mamumuntol]) = .000002

Zuraw (2000)

Using these to reason about acceptability of novel forms

(see p. 99, example (69))

Zuraw (2000)

Summary

- Probability of producing a SR given a particular UR depends on stochastic ranking
- Probability of attributing an incoming SR to a particular UR depends on probability of that UR, and the probably that the UR would yield the same SR in your own grammar
- Acceptability of novel items depends on the probability that you can associate it with a UR that would yield given SR

Zuraw (2000)

How does this compare with Boersma's proposal for evaluating competition with losing outputs?
How does it relate to Everett & Berent's comparative optimality proposal?

What these proposals have in common

- Both Zuraw's proposal and Everett & Berent's proposal rely on the grammar producing forms which "you wish it wouldn't"

- Faithfulness must be high enough to allow all patterns to emerge, but some fare worse on lower-ranked (subterranean) markedness constraints

- Differ on how those operate
  - ○ Everett & Berent: they simply assign their marks, and the relevant output suffers proportionately
  - ○ Zuraw: listener has to reason about probability that things could have gone differently

Switching gears: error rate estimation

$$\textit{apparent error rate} = \frac{\text{Errors on training sample}}{\text{Number of items in training sample}}$$

Switching gears: error rate estimation

$$\textit{apparent error rate} = \frac{\text{Errors on training sample}}{\text{Number of items in training sample}}$$

- Weiss & Kulikowski, p.24: “With an unlimited design sample used for learning, the apparent error rate will itself become the true error rate eventually. However, in the real world, we usually have relatively modest sample sizes with which to design a classifier and extrapolate its performance to new cases. For most types of classifiers, the apparent error rate is a poor estimator of future performance.. In general, apparent error rates tend to be biased optimistically. The true error rate is almost invariably higher than the apparent error rate.”

Error rate estimation

Two sources of inaccuracy in estimating error rate from the training set (reclassification):

- Training samples are finite
  - ○ Sample might be too small, and may differ from the true error rate simply because of probability
  - ○ (The sample wasn't truly representative)
- Hypothesis learned from the sample was too short-sighted
  - ○ Captures the training data well, but won't extend correctly to further examples
  - ○ *Overfitting*, or *overspecialization*
  - ○ What might an "overspecialized" solution be in phonology?

Error rate estimation

Dealing with error from uncertainty due to small sample size: confidence intervals

Image removed due to copyright considerations.

Weiss & Kulikowski's rule of thumb: by the time the sample size reaches 1000, the estimate is "extremely accurate" (By 5000, it's essentially equal to the true rate)

Error rate estimation

Dealing with short-sighted hypotheses (overfitting)

- Separate training data vs. test data ("holdout")
  - ○ Weiss & Kulikowski suggest convention of 2/3 training to 1/3 test
  - ○ Target size of testing set is 1000 or greater (why?)
  - ○ Proportion therefore must vary according to how much data is available

- Cross-validation
  - ○ "Leave-one-out"
  - ○ *k*-fold cross validation (*k* usually = 10)

Error rate estimation

Weiss & Kulikowski's general purpose suggestions:

- For $n > 100$, use cross validation (10-fold or leave-one-out)
- For $n < 100$, use leave-one-out
- For $n < 50$, try repeated 2-fold CV, or the .632 bootstrap:
- Take a sample of n items with resampling
- Train on that sample, test on anything not gotten in the sample to calculate error rate (e0)

  - ○ (On avg, that will lead to .632 samples chosen, .368 in test batch)

- e0 can also be approximated by repeated 2-fold cross validation (for reasons that are not clear to me)

- .632B = .368*e_{app} + .632*e0, where e_{app} = apparent error rate when trained on all cases

(Other than being quite complicated, why would we not want to do things like this for phonology data sets with less than 50 items?)

Error rate estimation

Back to Weiss & Kulikowski, p.24: “With an unlimited design sample used for learning, the apparent error rate will itself become the true error rate eventually. However, in the real world, we usually have relatively modest sample sizes with which to design a classifier and extrapolate its performance to new cases. For most types of classifiers, the apparent error rate is a poor estimator of future performance.. In general, apparent error rates tend to be biased optimistically. The true error rate is almost invariably higher than the apparent error rate.”

- Is it always the case that the apparent error rate will become the true error rate with an unlimited design sample? When might it not be?
  - May depend on what we mean by "true"

- When might resubstitution give a HIGHER error rate than cross-validation?

Error rate estimation

Weiss & Kulikowski, p. 46: Common mistakes

- Testing on the training data
  - "This mistake occurs infrequently these days, except perhaps for overzealous money managers promoting their successfully backfitted investment scheme"
- "Estimates for small sample sizes are much less reliable than those for large samples. A common oversight is that while the overall sample size may not be considered small, the subsample for any given class may be small. If the error rate on that class is particularly important, then the sample must be analyzed as a small sample."

Error rate estimation

Stepping back a minute:

☞ Why do all of these techniques seem sort of inapplicable to the task of learning a phonological grammar?

- What might we learn from them, even if we don't intend to use them directly?