Class 9: Production probability vs. acceptability

1 Why talk about acceptability?

- It seems pretty obvious that we want a grammar to attach a production probability to each candidate
  - After all, producing a candidate is something we have to do every time we talk; the grammar should model this
  - We can compare the grammar's production predictions to corpus or experimental data.
- But why acceptability ratings? Where do they come in?

2 Methodological reasons

- A lot of experiments ask subjects to rate forms
  - If we want to use those ratings to compare different grammar models (e.g., with and without some constraint), then we need the grammar to somehow output acceptability ratings
- Ratings as a proxy for unavailable frequency data
  - Temkin Martínez (2010) asked Hebrew speakers to rate a certain pronunciation of a real word
  - The probability that subjects gave a high rating to that pronunciation was taken as the probability that they would produce it (for purposes of fitting a grammar).

3 Theoretical reasons

- There are some real-life tasks that are kind of like acceptability rating
  - (unconsciously) deciding "could that have been a realization of *butter*?"
  - Perhaps deciding whether you like a new word well enough to use it, whether a rhyme is good enough for an improvised poem, whether a portmanteau blend is good enough to coin (*tofutastic?*)
  - Perhaps the competition between synonyms in production

Processes in the mind		*Things we can observe*
Grammar attaches production probability to each candidate	⟶	corpus frequency production frequency in experiment
Grammar attaches goodness rating to a form (in absolute terms, not just relative to other candidates)	? ⟶	acceptability ratings

4 Plan for today

- First, a sampling of empirical findings
- Second, a sample of modelling attempts
- Warning: there's not much out there in either set! But, we'll be able to draw some general conclusions

EMPIRICAL FINDINGS

5 An experiment gathering both production and rating data: Albright & Hayes 2003

- Past tense of nonce English verbs

- Production task:

(18)	**Screen:**	**Headphone input:**
Sentence 1	I dream that one day I'll be able to ___.	"I dream that one day I'll be able to *rife*."
Sentence 2	The chance to ___ would be very exciting.	"The chance to *rife* would be very exciting."
	Screen:	**Participant reads:**
Sentence 3	I think I'd really enjoy ___.	"I think I'd really enjoy [*response*]."
Sentence 4	My friend Sam ___ once, and he loved it.	"My friend Sam [*response*] once, and he loved it."

(p. 21)

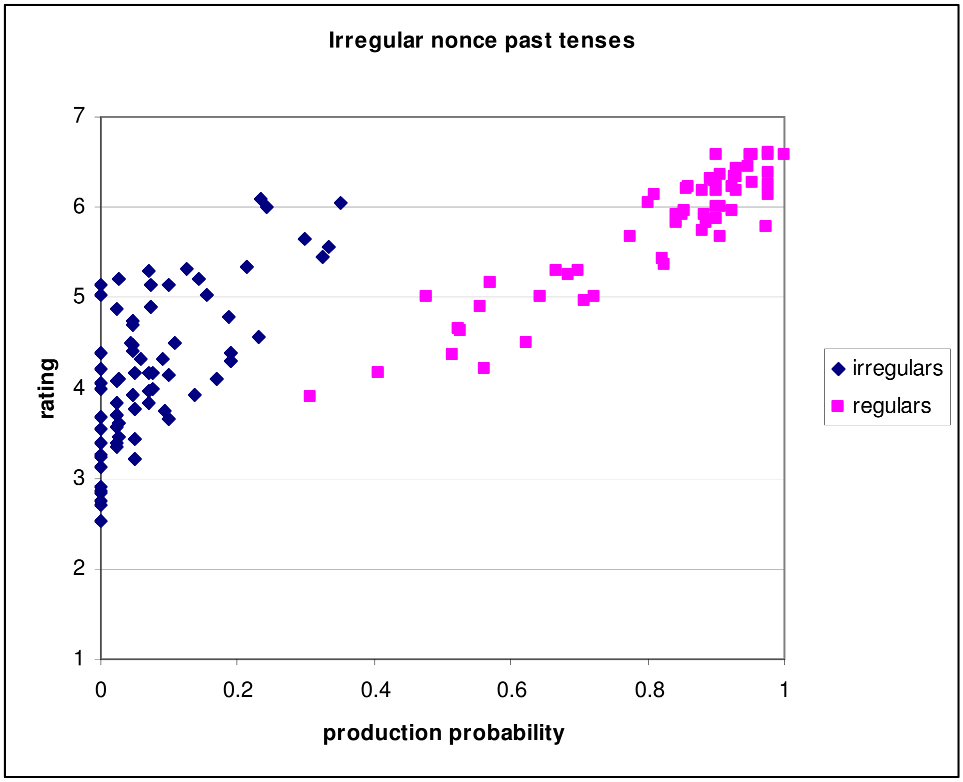
- Rating task:

(19) *Frame dialog for ratings task*

Sentence 1:	[voice]	"I dream that one day I'll be able to *rife*."
Sentence 2:	[voice]	"The chance to *rife* would be very exciting."
Sentence 3:	[participant]	"I think I'd really enjoy _____."
Sentence 4:	[participant]	"My friend Sam ______ once, and he loved it."
Sentence 5:	[voice]	"I dream that one day I'll be able to *rife*. My friend Sam *rifed* once, and he loved it."
	(participant rates)	
Sentence 6:	[voice]	"I dream that one day I'll be able to *rife*. My friend Sam *rofe* once, and he loved it."
	(participant rates)	

(p. 22)

- The authors kindly reported (and posted!) their results in detail, so we can make a plot:

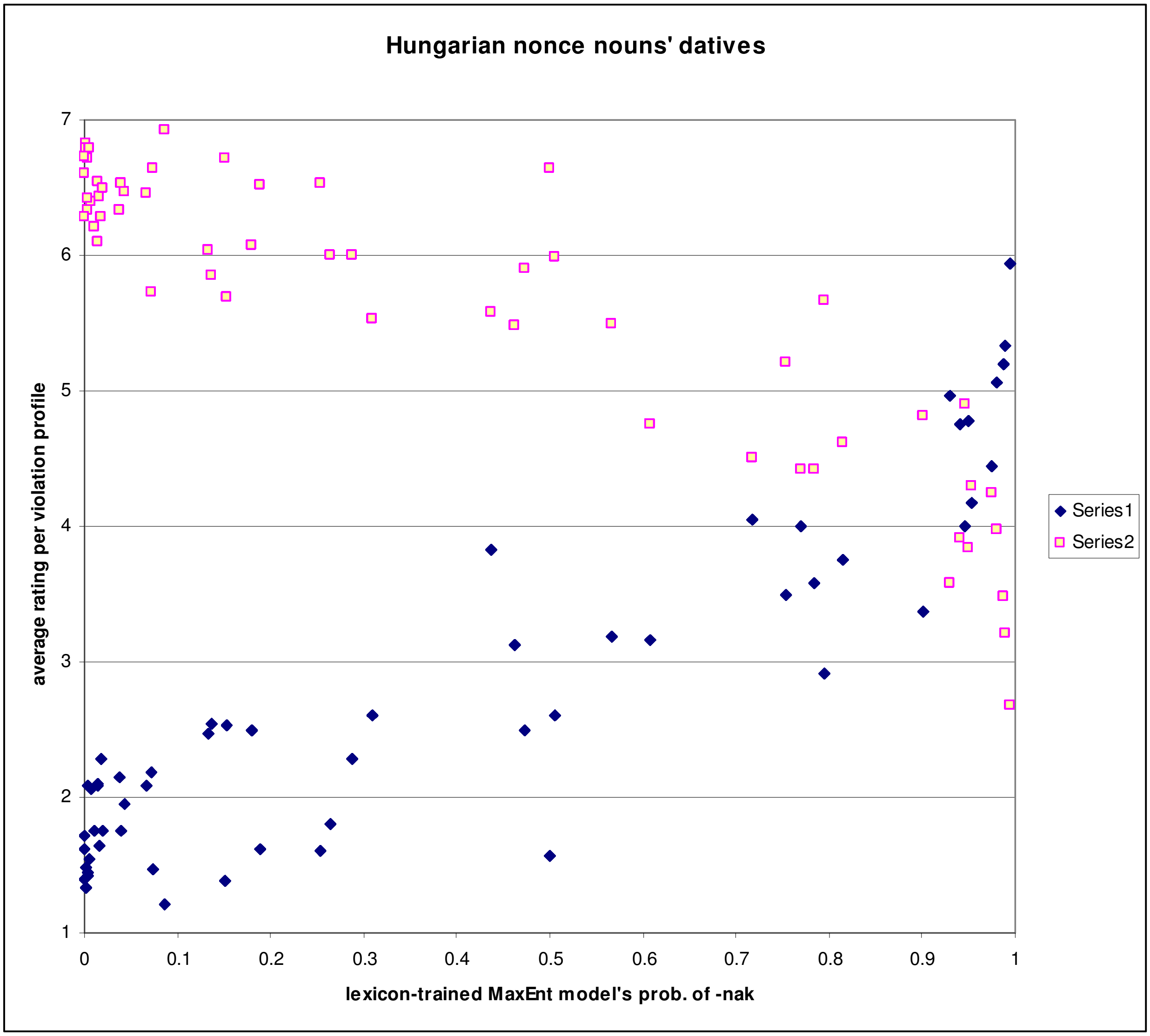


  - Regulars: nice, linear relationship
  - Irregulars: many items with 0 probability but a range of ratings; correlation less tight
  - Steeper slope for irregulars: small difference in probability → big difference in ratings
  - Let's discuss!

- Their model: each candidate has a "confidence" score based on the accuracy and scope of the best rule that generates it
  - e.g., for *gleed –gleeded*, best rule is Ø → əd / [X {d,t}__]$_{[+past]}$
- To generate predicted ratings, scale the set of these scores to have same mean and standard deviation as subjects' ratings.
- Production probabilities weren't explicitly generated, but paper does look at correlation with model's output—assumes a linearish relationship.

6 A paper with both frequencies and ratings: Hayes & Londe 2006; Hayes et al. 2009

- Hungarian speakers were asked to rate two options for the dative of a wug word: *-nak* and –*nek*.
- In this case, no true production probability available, but we can look at probability (of –*nak*) predicted by a MaxEnt model trained on the real lexicon (which is very accurate).
- Ratings are averaged over all items sharing a violation profile (only violation profiles with at least 10 items):



- The relationship does look linearish.

7 Temkin Martínez 2010: a fascinating study of mixed (lexical + free) variation

- We'll look just at the slice of the results where we can compare frequency and ratings
- Background—Hebrew spirantization: /p,b,k/ become fricatives / V__:

Consonant Pair	Root	Past	Infinitive or Future	Gloss
/p/ → [f]	/prs/	[paras]	[lifros]	*'spread'*
	/spr/	[safar]	[lispor]	*'count'*
	/nʃp/	[naʃaf]	[linʃof]	*'exhale'*
/b/ → [v]	/bnh/	[bana]	[livnot]	*'build'*
	/sbl/	[saval]	[lisbol]	*'suffer'*
	/gnb/	[ganav]	[lignov]	*'steal'*
/k/ → [χ]	/ktb/	[katav]	[liχtov]	*'write'*
	/mkr/	[maχar]	[limkor]	*'sell'*
	/drk/	[daraχ]	[lidroχ]	*'step'*

(p. 23)

- But! There are exceptional always-stops (from Tiberian Hebrew non-alternating stops that neutralized with *p,b,k*), and exceptional always-fricatives (from Tiberian Hebrew non-alternating continuants that neutralized with *f,v,χ*):

		Root	3rd Person Sg. Past	Infinitive	
a.	/k/ (< *k)	/ktb/	[katav]	[liχtov]	'to write'
b.	/k/ (< *q)	/krʔ/	[kara]	[likro]	'to read'

(p. 28)

		Root	3rd Person Sg. Past	Infinitive	
a.	/v/ (<* w)	/vtr/	[viter]	[levater]	'to give up'
	/χ/ (<* ħ)	/χps/	[χipes]	[leχapes]	'to look for'
b.	[v] (<* b)	/btl/	[bitel]	[levatel]	'to cancel'
	[χ] (<* k)	/kpr/	[kiper]	[leχaper]	'to atone'

(p. 29)

- (There can even be a mix of these within a word:

Root	3rd Person Sg. Past	Infinitive	
/bkr/	[biker]	[levaker]	'to visit
/kbr/	[kavar]	[likbor]	'to bury'

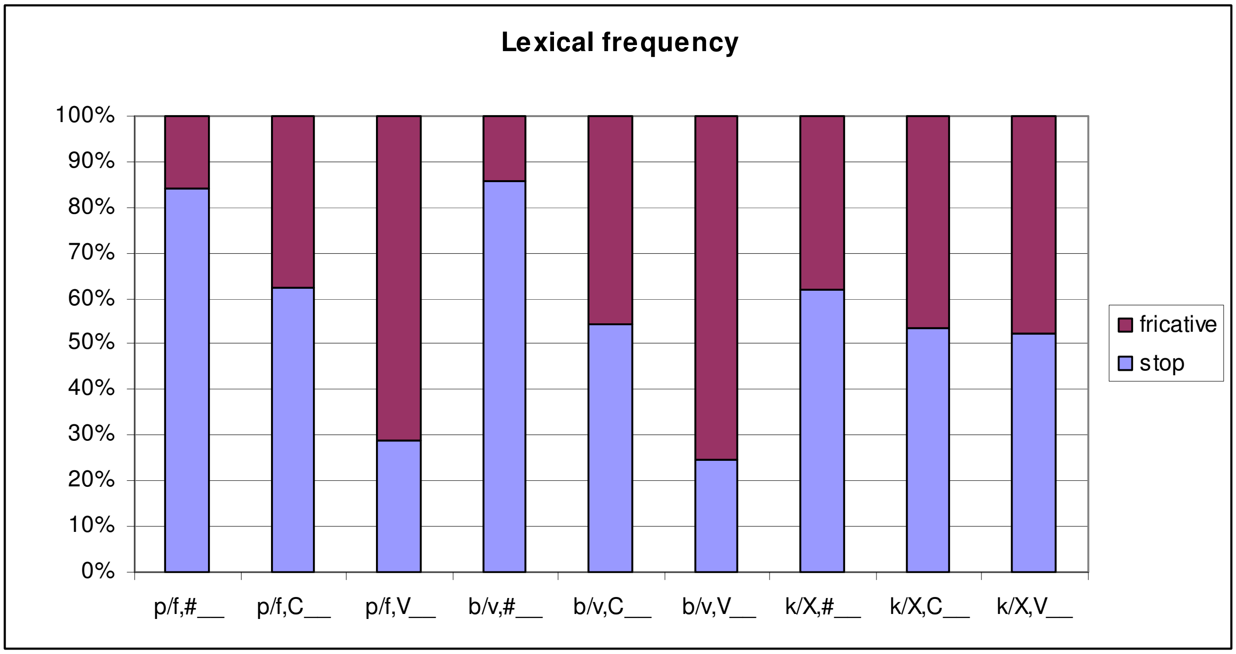
(p. 7))

- Perhaps because of this lexical variation, there's also some free variation:

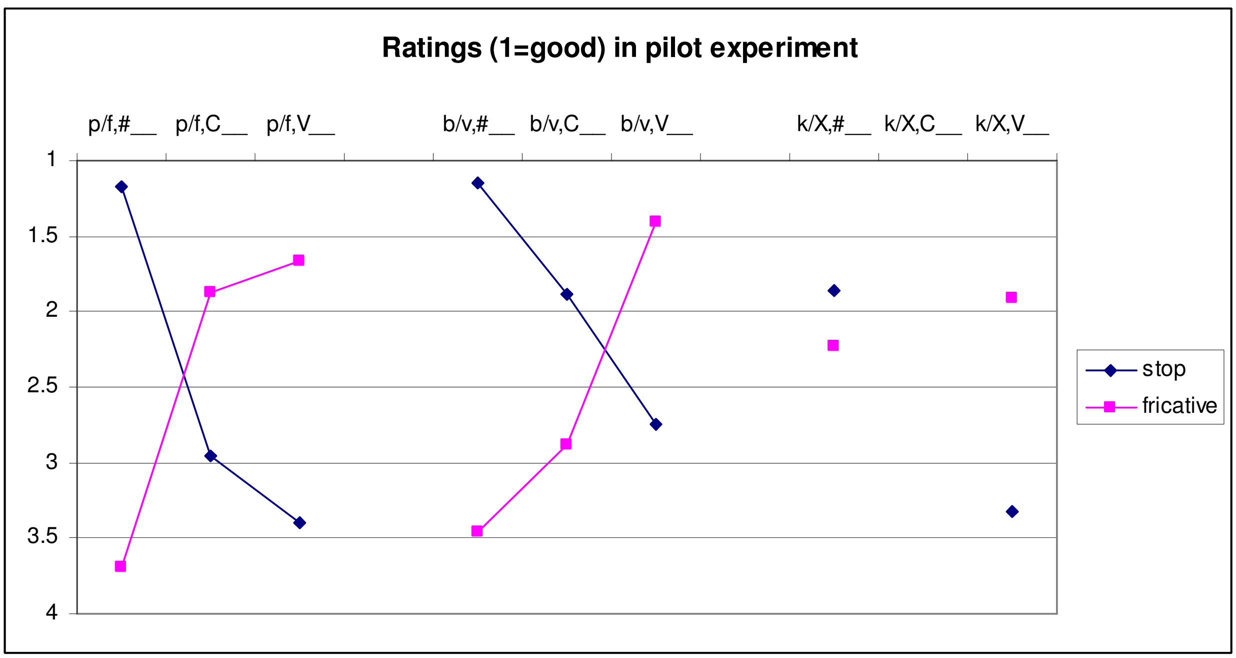
Expected	Acceptable Variant	Gloss
pagaʃ	fagaʃ	'met'
jikbor	jikvor	'will bury'
jeχase	jekase	'will cover'

(p. 8)

- Lexical statistics, based on the LLHN (Bolozky & Becker 2006), singular nouns only



- To see how widespread this variation is, Temkin Martínez had Hebrew speakers rate 2 pronunciations *of a real word*:



  - Just as in lexical statistics, C__ tolerates spirantization better than #__ does
  - Also as in lexical statistics, #fricative is more tolerated in velars than in labials.

MODELS? (there's not that much out there)

8 Boersma & Hayes 2001 (as you read): sigmoid relationship

- Hayes (1997) had gathered ratings of English light and dark /l/ in different contexts.
- To test the Gradual Learning Algorithm, they needed to convert these into candidate probabilities.
- Call *darkRating – lightRating* ΔJ
- Predicted probability of light-*l* candidate = $\frac{1}{1+0.2^{\Delta J}}$, where 0.2 was probably hand-fitted and would presumably depend on the range of the rating scale subjects use.
- Conversely, predicted $\Delta J = \frac{\log\left(\frac{1}{probOfLight} - 1\right)}{\log 0.2}$.

Word type	Judged as light	Judged as dark	Judgment Difference	Conjectured Frequency of Light Variant
a. *light*	1.30	6.10	4.80	99.956%
b. *Louanne*	1.10	5.55	4.45	99.923%
c. *gray-ling, gai-ly, free-ly*	1.57	3.34	1.77	94.53%
d. *Mailer, Hayley, Greeley, Daley*	1.90	2.64	0.74	76.69%
e. *mail-er, hail-y, gale-y, feel-y*	3.01	2.01	–1.00	16.67%
f. *mail it*	4.40	1.10	–3.30	0.49%
g. *bell, help*	6.60	1.12	–5.48	0.0011%

(p. 32 of ms.)

9 Boersma 2005 adds a twist: perception grammar

- The "prototypicality" problem: if you ask listeners to pick the best instance of [i], they'll tend to choose one that's very high and front, even though this isn't the most frequent realization.
  - To view this as a rating issue, imagine that the subject is asked to rate each token rather than just pick the best one
- Boersma's solution: run the form through the perception grammar

[380 Hz] (UF = |i|)	320 Hz not /a/	380 Hz not /a/	460 Hz not /i/	320 Hz not /e/	460 Hz not /a/	380 Hz not /i/	380 Hz not /e/	320 Hz not /i/	460 Hz not /e/
/a/		*!							
☞ /e/							←*		
√ /i/						*!→			

(p. 7 of ms.)

  - The more probable the /i/ candidate, the better the stimulus is judged (I might be reading in here).

o Let's discuss how this would apply to something more phonological, like, say the /l/ ratings above.

10 Albright (n.d.): acceptability as probability of being well-formed

- This paper is about phonotactics. so, we can think of "variation" as being in the degree to which words of the given type exist.
- Builds on Bailey & Hahn 2001's Generalized Neighborhood Model (itself adapted from Nosofsky's Generalized Context Model)

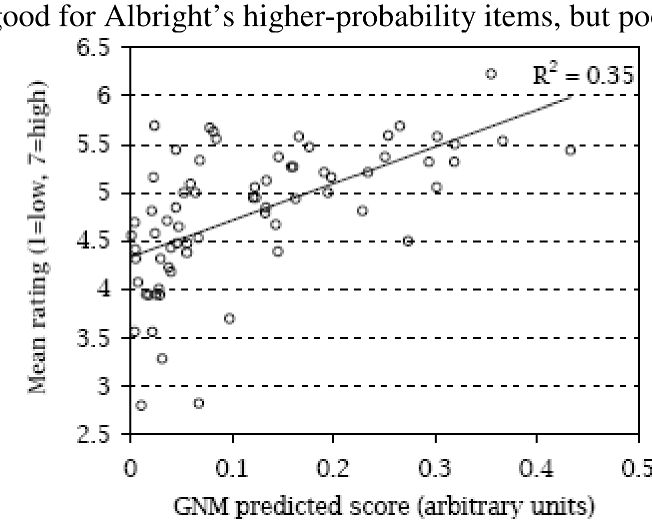
- If evaluating a potential word *i*, determine the probability that it belongs to the set "English"

$$\text{probability}(plake \in \text{English}) \propto \sum_{c \in English} FrequencyWeightedSimilarity(plake, c)$$

- How do we get similarity of *plake* and, say, *bake*? $e^{(-d_{plake,bake}/s)^P}$
  - where $d_{plake,bake}$ is the string-edit distance between *plake* and *bake*
  - *s* and *P* are free parameters—Albright uses 0.1739 and 1.
  - $d_{plake,\ bake}$ = 1.4 (1 insertion, 1 deletion, penalty of 0.7 for each—Albright does something more subtle, taking advantage of similarity of *p* and *b*)
  - So, similarity(*plake*, *bake*): $e^{(-1.4/0.1739)}$ = 0.000319

- Multiply by CELEX frequency of *bake*: 423 * 0.000319 = 0.134899
- Repeat the procedure for every other word of English, sum up the frequency-weighted similarities.
  - The result (in arbitrary units) should be proportional to probability (from listener's point of view) that it's an English word.

- Albright's idea is that ratings should be a (linear?) function of this value.
- It's pretty good for Albright's higher-probability items, but poor for lower-probability:



Predicted scores vs. Albright's subjects' ratings (70 random filler items).

(Albright, p. 9)

- Albright actually argues for something quite different instead, based on extracting and attaching numbers to strings of natural classes (based on frequency and successful specificity)—but this isn't a course about modelling phonotactic probability!

11 Becker & Gouskova 2012: consider the "sub-grammars" a word could belong to

- "Yer" study: asks Russian speakers to accept or reject suffixed nonce words with and without mid-V deletion

> In this river, there lives a long <u>şer</u>
>
> Rate the underlined word. Can it be a word of Russian?
> it cannot [1] [2] [3] [4] [5] it can
>
> Ivan caught a long <u>şra</u>
> Can this word be a declined variant of the word şer?
> [no] [yes]
>
> Ivan caught a long <u>şera</u>
> Can this word be a declined variant of the word şer?
> [no] [yes]

(p. 11)

- The paper's goal is to model the no/yes responses.
- Assumes that words belong to different subgrammars. W.r.t. forming the genitive,
  - Add [-a] (non-yer masculines)
  - Add [-a] and delete stem's last V (yer masculines)
  - *something else for feminines*
  - *something else for neuters*
- Could be literally different rankings, or could be lexical indexation of constraints
- Each of these sub-grammars has two parts
  - a phonotactic grammar (learned using Hayes & Wilson 2006)—tells you how good a word is as a member of that sub-lexicon
  - an input-output-mapping grammar—here, forms genitive from nominative (assumed to be UR)

- The model will have these ingredients:
  - PhonotacticGrammar$_{yer}$, MappingGrammar$_{yer}$, PhonotacticGrammar$_{non\text{-}yer}$, etc.

- To do the experimental task...
  - the speaker sums the probabilities that the proposed mapping gets under all the groups
  - weighted by how probable it is that the word belongs to that group
  - (for simplicity, we'll ignore the feminine and neuter groups)

p(şer ~ şra) = p(*yer*| şer) * p(şra|*yer*) + p(*non_yer*| şer)*p(şra|*non_yer*)

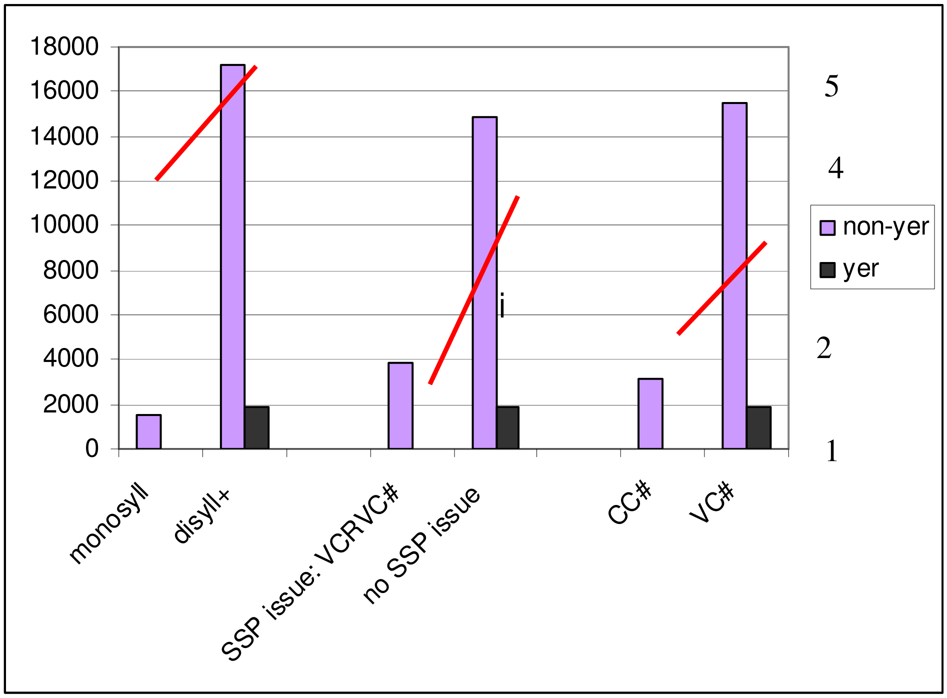
pretty much 0

p(şer|PhonotacticGrammar$_{yer}$)*p(*yer*) / p(şer)

~10% of words belong to yer group

p(şer|PhonotacticGrammar$_{yer}$) +
p(şer|PhonotacticGrammar$_{nonyer}$)
estimated through ratings of
nominative (1-5 scale)

- What I couldn't find in the paper was a comparison of this model's predictions to the actual ratings.
- But, in Gouskova & Becker to appear there are some similar data where we can at least compare V-deletion probability in real words to ratings in similar wug words:



  - Lexical data: V deletion ("yer") is almost forbidden...
    - in monosyllables (lóp, lb-óf)
    - if V-deletion creates a medial Sonority Sequencing Principle violation (ágn^{j}its , ágnts-əf)
    - if the stem ends in CC (hypothetical pést, pst-óf)
  - The overlain lines are my attempt to add the median rating for V-deleted (yer) items in that group (scale on right)

- Let's discuss this idea of sub-grammar assignment for the cases we've seen so far.

References

Albright, Adam. Natural classes are not enough: biased generalization in novel onset clusters. Manuscript. MIT, ms.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90(2). 119–161. doi:10.1016/S0010-0277(03)00146-X (29 April, 2013).

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language* 44(4). 568–591. doi:10.1006/jmla.2000.2756 (28 April, 2013).

Becker, Michael & Maria Gouskova. 2012. Source-oriented generalizations as grammar inference in Russian vowel deletion. Manuscript. Indiana University and New York University, ms.

Boersma, Paul. 2004. A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments.

Boersma, Paul. 2005. *Prototypicality judgments as inverted perception*. University of Amsterdam.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32. 45–86.

Bolozky, Shmuel & Michael Becker. 2006. Living Lexicon of Hebrew Nouns. University of Massachusetts, Amherst, ms. http://becker.phonologist.org/LLHN.

Gouskova, Maria & Michael Becker. to appear. Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language and Linguistic Theory*.

Hayes, Bruce. 1997. *Gradient well-formedness in Optimality Theory*.

Hayes, Bruce & Zsuzsa Cziráky Londe. 2006. Stochastic Phonological Knowledge: The Case of Hungarian Vowel Harmony. *Phonology* 23(01). 59–104. doi:10.1017/S0952675706000765.

Hayes, Bruce, Péter Siptár, Kie Zuraw & Zsuzsa Londe. 2009. Natural and Unnatural Constraints in Hungarian Vowel Harmony. *Language* 85(4). 822–863. (13 February, 2011).

Hayes, Bruce & Colin Wilson. 2006. A Maximum Entropy Model of Phonotactics and Phonotactic Learning.

Temkin Martínez, Michal. 2010. Sources of non-conformity in phonology: variation and exceptionality in Modern Hebrew spirantization. University of Southern California Ph.D. Dissertation.