Class 3 (9 April 2013): Probability distributions over Classic OT grammars

To do
- Reading: Coetzee 2009
- Software exercise (see instructions on web page): Playing with the GLA

Overview: Variation can be conceptualized as vacillation between grammars. "Partial ranking" and Stochastic OT take this approach, but put restrictions on probability distributions over Classic OT grammars.

1 What is a probability distribution?

- It's a function from possible outcomes (of some random variable) to probabilities.
- Some well-known examples
  - Flipping a fair coin

which side lands up	*probability*
head	0.5
tails	0.5

  - Rolling 2 fair dice and summing the numbers

sum of number of dots on top faces	*probability*
1	0.00
2	1/36 = 0.03
3	2/36 = 0.06
4	3/36 = 0.08
5	4/36 = 0.11
6	5/36 = 0.14
7	6/36 = 0.17
8	5/36 = 0.14
9	4/36 = 0.11
10	3/36 = 0.08
11	2/36 = 0.06
12	1/36 = 0.03

2 Probability distributions over grammars

- One way to think about within-speaker variation is that, at each moment, the speaker has multiple grammars to choose between.

set of phrase structure rules	*probability*
{S → NP VP, VP → V NP }	0.4
{S → VP NP, VP → NP V}	0.6

  ⇒ 40% John washed the dog, 60% John the dog washed

  - This is sort of the view in Niyogi 2009—where the variation is mostly across speakers, not within a speaker.

- Typically, the probability distribution is constrained somehow, by attaching probabilities not to whole grammars but to parameters of some kind

  - Putting a probabilistic spin on a proposal of Haegeman 1987:

settings of 2 binary parameters	*impossible set of probabilities*	*possible set of probabilities*
determiner-head order, object drop	0.4	0.04
determiner-head order, no object drop	0.1	0.36
head-determiner order, object drop	0.2	0.06
head-determiner order, no object drop	0.3	0.54

If determiner-head, object drop more probable; if head-determine, no object drop more probable.

  - Why is the possible set possible?
    - Because it can be derived by attaching one probability to each binary parameter
    - I.e., what speakers really represent is a probability distribution <u>over each parameter</u>, which gives rise epiphenomenally to the overall probability distribution:

settings of order parameter	*probability*
determiner-head order	0.4
head-determiner order	0.6

settings of object-drop parameter	*probability*	*example (Haegeman 1987)*
object drop	0.1	Beat till elastic
no object drop	0.9	Beat the mixture till elastic

  - Related proposal in Adger & Smith 2010: the grammar itself doesn't vary, but important lexical entries can:

lexical entry of T[past]	*probability*
T[tense:past, *u*num:, *u*pers:]	0.33
T2[tense:past, *u*case:nom, *u*pers:]	0.67

  ⇒ 33% we were, 67% we was (Buckie dialect of British English)

3 Probability distributions over Classic OT grammars

- In the general case, they might look like this:

ranking	*probability*	*example output*
MAX-C >> *θ >> IDENT(continuant)	0.10	t̪ɪn
MAX-C >> IDENT(continuant) >> *θ	0.50	θɪn
*θ >> MAX-C >> IDENT(continuant)	0.05	t̪ɪn
*θ >> IDENT(continuant)>> MAX-C	0.20	ɪn
IDENT(continuant) >> MAX-C >> *θ	0.05	θɪn
IDENT(continuant) >> *θ >> MAX-C	0	ɪn

  - But I haven't seen any proposal along those lines
    - One possible challenge : does the child have to learn *n*!-1 probabilities, if *n* constraints ?
  - Instead, the probability distributions are constrained

4 Putting restrictions on probability distributions—let's brainstorm

- Maybe instead the speaker learns a number for each constraint, and derives the probability distribution from that
- Or the speaker learns a number for each pair of constraints
- or...

5 Back to variable constraint ranking

- Recall that for Labov's New York City (th) case, we could say this:

/θɪk/	*θ	IDENT(cont)
☞ *a* [θɪk]	*	
☞ *b* [t̪ɪk]		*

Jagged line: constraint ranking *varies*.

Dotted line: ranking is *unknown*.

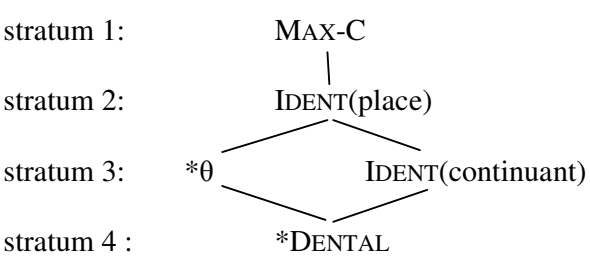
- More of the ranking:

/θɪk/	MAX-C	IDENT(place)	*θ	IDENT(cont)	*DENTAL
☞ *a* [θɪk]			*		*
☞ *b* [t̪ɪk]				*	*
c [ɪk]	*!				
d [sɪk]		*!			

- This is like saying:
  - MAX-C >> IDENT(place) >> *θ >> IDENT(continuant) >> *DENTAL: something >0 %
    - and/or IDENT(place) >> MAX-C >> *θ >> IDENT(cont) >> *DENTAL: something >0 %
  - MAX-C >> IDENT(place) >> IDENT(continuant) >> *θ >> *DENTAL: something >0 %
    - and/or IDENT(place) >> MAX-C >> IDENT(cont) >> *θ >> *DENTAL: something >0 %
  - all other rankings : 0%

6 Anttila's proposal

- A speaker's grammar is a ranking of constraints into *strata*:



for convenience, I'll assume that MAX-C >> IDENT(place), although we can't know

- If a stratum has more than one constraint in it, every time the speaker makes a tableau she/he must arrange those constraints into a linear order
  - each order of constraints within a stratum is <u>equally probable</u>
- In our example, this means we have the following probability distribution over rankings:
  - 50% : MAX-C >> IDENT(place) >> *θ >> IDENT(continuant) >> *DENTAL
  - 50% : MAX-C >> IDENT(place) >> IDENT(continuant) >> *θ >> *DENTAL
  - 0% : all other rankings
- Let's discuss the restrictions that this theory places on probability distributions: what are some things that can't happen?

7 Finnish example

- Anttila proposes the following constraint ranking for Finnish genitives:

(50) The grammar for Finnish, final version

Set 1	Set 2	Set 3	Set 4	Set 5
*X.X	*L	*H/I	*H/O	*H/A
	*H	*Í	*Ó	*Á
		*L.L	*L/A	*L/O ≫ *L/I
			*H.H	*A ≫ *O ≫ *I
			*Ĥ	*L
			*X.X	

(Anttila 1997, p. 21)

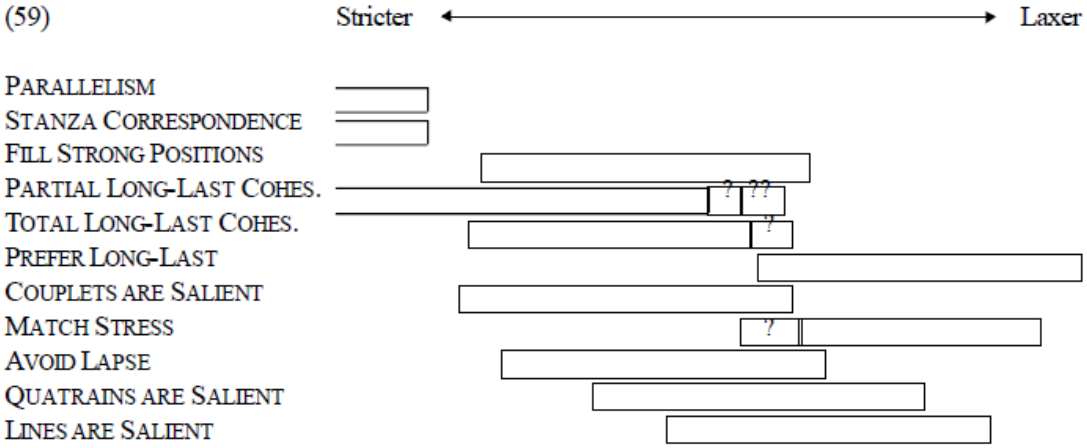
- The predicted outcomes depend on the set of constraints
  - This is true in any type of model
  - But it's even more true here because the theory puts such strong restrictions on the probability distribution over rankings
  - In particular, you can never get, e.g., a 90%-10% distribution with just 2 constraints.

A THEORY THAT'S A BIT LESS RESTRICTIVE: STOCHASTIC OT

8 The basic idea

- Assign each constraint to a range on the number line.
  - Early version of the idea from Hayes & MacEachern 1998. Each constraint is associated with a range, and those ranges also have fringes, indicated by "?" or "??"

(59)



(Hayes & MacEachern 1998, p. 43)

- Each time you want to generate an output, choose one point from each constraint's range, then use a total ranking according to those points.

- Discuss this claim: This approach defines (though without precise quantification) a probability distribution over constraint rankings.

9 Stochastic OT (Boersma 1997; Boersma & Hayes 2001)

- This was the first theory to quantify ranking preference.
- "stochastic" just means "probabilistic", so various theories could be described as "stochastic OT". With a capital *S*, though, I mean specifically this theory

- In the grammar, each constraint has a "ranking value":
  *θ 101
  IDENT(cont) 99
- Every time a person speaks, they add a little noise to each of these numbers, then rank the constraints according to these numbers.

⇒ Go to demo [03_StochOT_Materials.xls]

- Discuss claim: once again, this defines a probability distribution over constraint rankings
- Discuss claim: an Anttilan grammar is a special case of a Stochastic OT grammar
- Researchers who use this model often acknowledge stylistic conditioning, but idealize away from it. Ideas on how we could modify the model to add in the effect of style?

10 The Gradual Learning Algorithm (Boersma 1997; Boersma & Hayes 2001)

- This was a groundbreaking aspect of the proposal: it came with a procedure for learning the values.
  - Important theoretically: if this is a theory of what a person's grammar looks like, we need some theory of how the grammar gets that way, during childhood and beyond
  - Important practically: it meant you could apply these models to your own data

Procedure:

1. Suppose you're a child. You start out with both constraints' ranking values at 100.
2. You hear an adult say something—suppose /θɪk/ → [t̪ɪk]
3. You use your current ranking values to produce an output. Suppose it's /θɪk/ → [θɪk].
4. Your grammar produced the wrong result! (If the result was right, repeat from Step 2)
5. Constraints that [t̪ɪk] violates are ranked too low; constraints that [θɪk] violates are too high.
6. So, promote and demote them, by some fixed amount (say 0.33 points)

/θɪk/	*θ	IDENT(cont)
the adult said this [θɪk]	* demote to 99.67	
your grammar produced this [t̪ɪk]		* promote to 100.33

7. Repeat.

⇒ Go to demo (same Excel file, different worksheet)

- Suppose, as in our demo, that adults produce [t̪ɪk] 90% of the time. Will your grammar ever stop making errors?
- What's the effect of the column (in Excel file) labeled 'plasticity'?
- What if the adults actually don't vary, and the outcome is always [θɪk]. What will happen to the ranking values? (After discussing, let's try it in the spreadsheet.)
- In that case, will your grammar ever stop making errors?

11 Software

- See this week's software exercise (on web page)

- OTSoft (Hayes & al. 2003, available from Bruce's web page)
  - user-friendly
  - Windows only
  - allows you to run GLA on a dataset, or test the behavior of a StOT grammar (including an Anttilan one)

- Praat (Boersma & Weenink 2006, available from www.praat.org)
  - harder to learn
  - platform-general

SOME STOCHASTIC OT/GLA CASE STUDIES

12 Albright & Hayes 2006: "Junk" constraints

- Albright & Hayes 2006 is one of a series of papers developing a model for learning constraints from morphological mappings.

Navajo sibilant harmony:

(3) a.	[bà:ʔ]	[sì-bà:ʔ]
b.	[č'ił]	[ši-č'ił]
c.	[č^{h}ò:jìn]	[ši-č^{h}ò:jìn]
d.	[gàn]	[sì-gàn]
e.	[k'àz]	[sì-k'àz]
f.	[kéšgã̀:]	[ši-kéšgã̀:], [sì-kéšgã̀:]
g.	[sí:ʔ]	[sì-sí:ʔ]
h.	[tã̀š]	[ši-tã̀š], [sì-tã̀š]
i.	[tı́]	[sì-tı́]
j.	[t͡ɬé:ž]	[ši-t͡ɬé:ž], [sì-t͡ɬé:ž]

(Albright & Hayes 2006, p. 3)

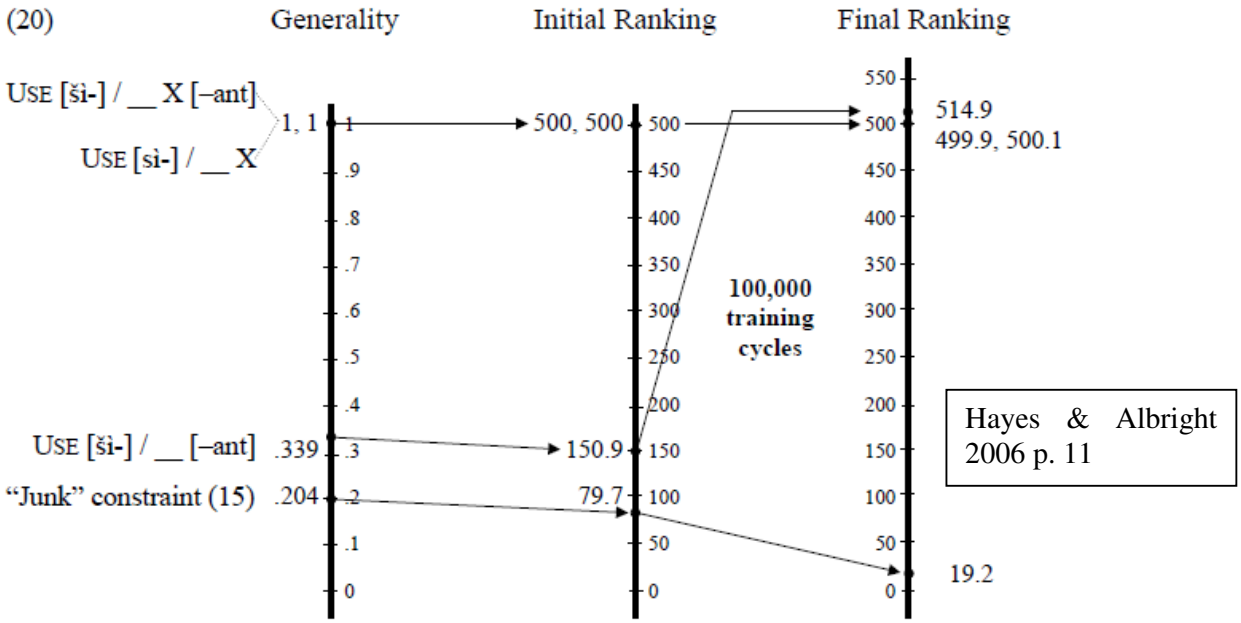
- The Albright/Hayes learner (which we won't get into) learns some sensible constraints like these:
  - USE ŠI / __[-anterior] local harmony: should be high-ranked
  - USE ŠI / __X[-anterior] distal harmony: should be mid-ranked
  - USE SI default is [+ant] should be lower-ranked
- but also some "junk" constraints like these:
  - USE SI / __([-round])* [+ant, +cont] ([-cons])* # happens to be true in training data but probably not high-ranked in real grammar

==> Demo: let's see what happens if we apply GLA to a schematic case like this [03_Navajo_for_GLA_revised.xls]

- Albright & Hayes's solution:
  - constraints' **initial ranking values** reflect their **generality**

▪ generality of USE ŠI/__[-anterior] = (# of __[-anterior] words) / (# of words that use ŠI) = 19/56 = 0.34
  - generality of junk constraint = 37/181 = 0.20
  - These numbers are then scaled so that they range from 0 to 500 (see paper for details):

(20)



- Why does it work?
  - USE SI ranks high enough from the beginning to avoid errors like /si+tala/ → [šitala]
  - So, the junk constraint never gets promoted
- Albright and Hayes would probably both favor a MaxEnt approach now
  - But this is a nice demonstration of how to introduce **bias** into a learner
    - Purpose in this case: prevent "overfitting" to arbitrary details of learning data
  - In this case, the desired bias is not to put too much trust into constraints that cover only few cases

13 Boersma & Levelt 2000: predicting acquisition order

- The G in GLA stands for "gradual"
  - The algorithm doesn't just return its final grammar
  - Instead, it gradually updates the initial grammar
  - At every step, it's possible to "pause" the grammar and ask what its current output is
  - This provides a concrete analogy to child language acquisition

- Levelt had previously done work on the order in which different syllable types are produced by children.
  - Data for 12 children acquiring Dutch:

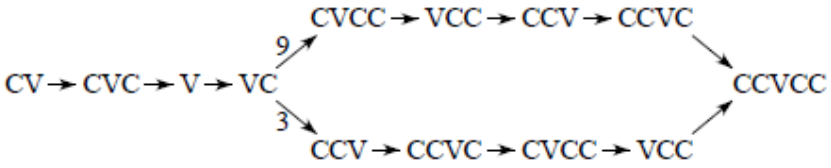


Figure 1. Acquisition order for syllable types in Dutch.

(Boersma & Levelt, p.1)

- Boersma & Levelt fed the GLA the frequencies of the faithful syllables in Dutch
  - They also, importantly, set markedness constraints' initial ranking value to 100, and faithfulness constraints' to 50
  - This means that initially, all outputs will be [CV], regardless of input
  - Gradually, FAITH climbs, markedness constraints fall, and other syllable shapes get produced

Boersma & Levelt's resulting ranking values over time

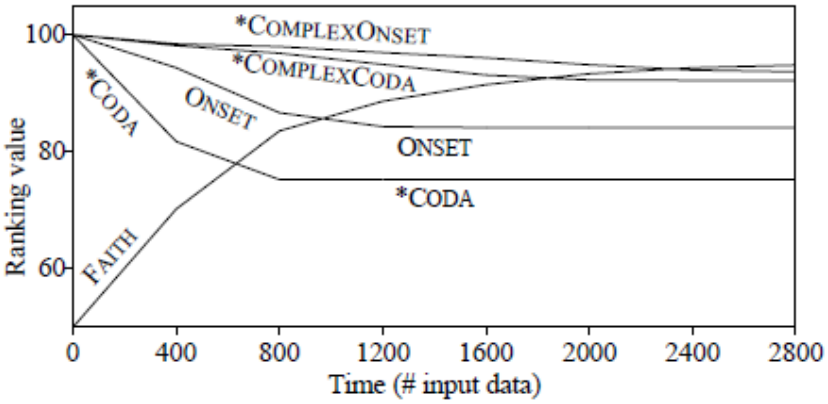


Figure 2. Constraint rankings as functions of time.

(Boersma & Levelt p. 5)

Here are the rates of correct (faithful) production for each syllable type over time

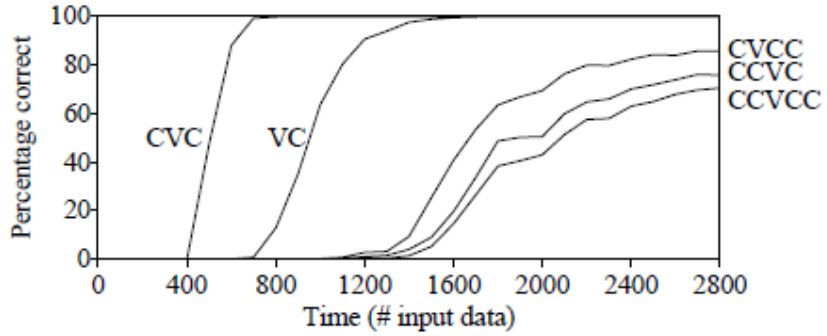


Figure 3. Five learning curves for our simulated learner.

(Boersma & Levelt p. 7)

- Jarosz 2010 shows how the frequencies of the different syllable types in the language matter
  - if a syllable type is rare, then errors on it are always rare
    - ...even if the accuracy of the current grammar on these syllable types is low
  - A markedness constraint's demotion rate depends on how many word tokens violate it

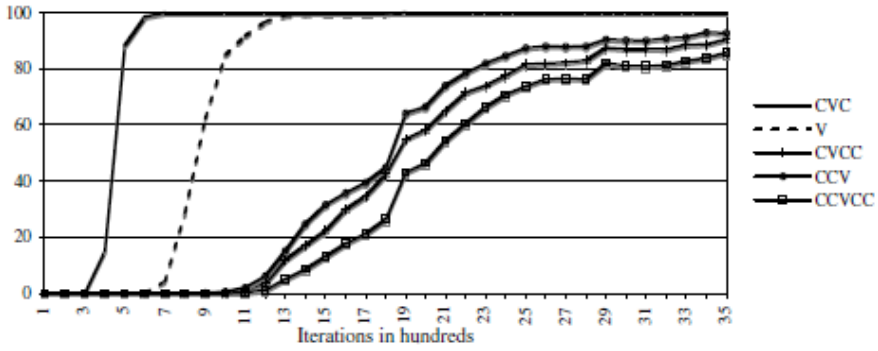
- Jarosz's results (following the Boersma & Levelt procedure) for Dutch, English, and Polish
  - Same constraints, faithful candidate is always the winner—but the input frequencies differ.

(results on next page)

TABLE 1. *Relative frequencies of syllable types in Dutch*

CV	CVC	CVCC	V	VC	VCC	CCV	CCVC	CCVCC
44·8%	32·1%	3·3%	3·9%	12·0%	0·4%	1·4%	2·0%	0·3%

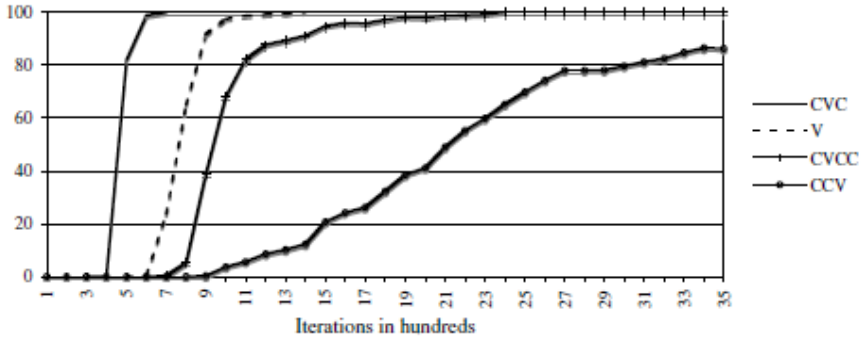
Jarosz p. 573



Dutch, Jarosz p. 594

TABLE 5. *Relative frequencies of syllable types in English*

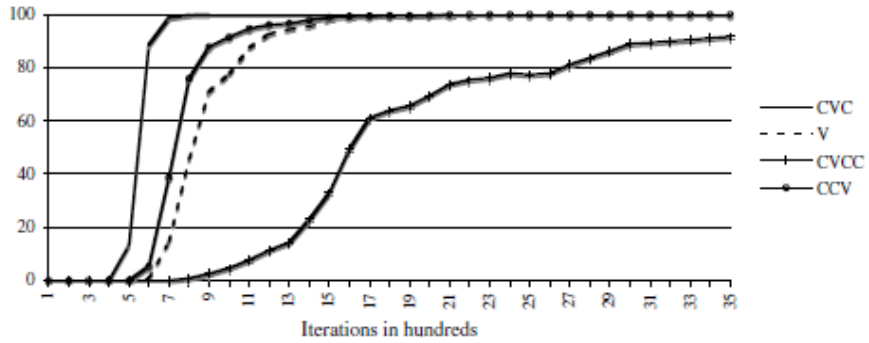
CV	CVC	CVCC	V	VC	VCC	CCV	CCVC	CCVCC
24·4%	40·5%	10·1%	4·7%	13·0%	3·5%	0·9%	2·2%	0·6%



Jarosz p. 598

TABLE 6. *Relative frequencies of syllable types in Polish*

CV	CVC	CVCC	V	VC	VCC	CCV	CCVC	CCVCC
50·3%	20·9%	3·3%	8·5%	3·5%	0·6%	9·6%	4·0%	0·6%

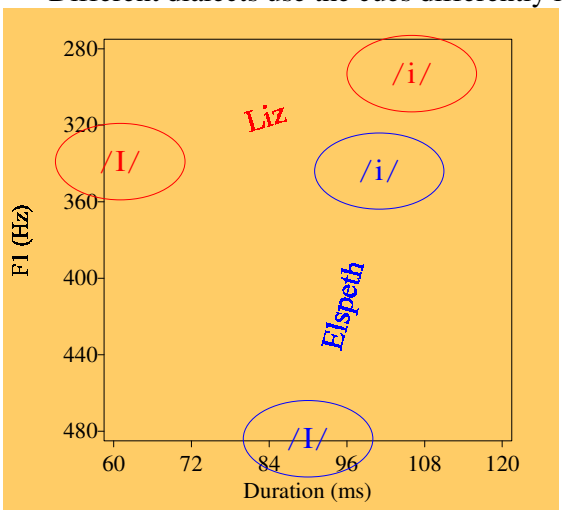


Jarosz p. 600

14 Escudero & Boersma 2001: learning to weight perceptual cues

- English /i/ ("peach") and /ɪ/ ("pitch") are differentiated by two main cues
  - F1 (≈ tongue/jaw lowering) and duration

- Different dialects use the cues differently in production:



"Elspeth": a Scottish English speaker
"Liz": a Southern British English speaker

Escudero & Boersma slide 6

  - So the boundary between the two categories in this 2-dimensional space is something that must be learned

- Boersma has been a proponent of using OT tableaux for perception. Example:

[350 Hz, 80 ms]	350 Hz ≠ /ɪ/	350 Hz ≠ /i/	80 ms ≠ /ɪ/	80 ms ≠ /i/
perceive as /i/		*		*
perceive as /ɪ/	*		*	

This is too low to be [ɪ] in Southern dialect, too short (and high) to be [i] in Scottish dialect

- ==> Let's step through Escudero & Boersma's results graphs on screen
  - training data for each dialect: typical realizations of [i] and [ɪ]
  - If learner's current grammar miscategorizes this item, ranking values are adjusted
  - At the end, one can test how often each Hz-msec combination is categorized as each vowel

15 Coming up

- Computational problems with GLA were probably a big factor in pushing phonologists towards other models
- Noisy Harmonic Grammar and MaxEnt OT: abandon strict constraint ranking for other ways of resolving constraint conflicts (weighting)

References

Adger, David & Jennifer Smith. 2010. Variation in agreement: A lexical feature-based approach. *Lingua* 120(5). 1109–1134. doi:10.1016/j.lingua.2008.05.007 (9 April, 2013).

Albright, Adam & Bruce Hayes. 2006. Modeling productivity with the Gradual Learning Algorithm: the problem of accidentally exceptionless generalizations.. In Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky, & Ralf Vogel (eds.), *Gradience in Grammar: generative perspectives*. Oxford: Oxford University Press.

Anttila, Arto. 1997. Deriving variation from grammar.. In Frans Hinskens, Roeland van Hout, & W. Leo Wetzels (eds.), *Variation, Change, and Phonological Theory*, 35–68. Amsterdam: John Benjamins.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21. 43–58.

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

Boersma, Paul & Clara C Levelt. 2000. Gradual constraint-ranking learning algorithm predicts acquisition order.. In Eve V Clark & Eve V Clark (eds.), *The Proceedings of the Thirtieth Annual Child Language Research Forum*. Stanford, CA: CSLI Publications.

Boersma, Paul & David Weenink. 2006. Praat: Doing phonetics by computer, version 4.4. http://www.praat.org/.

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.

Coetzee, Andries W & Joe Pater. 2011. The place of variation in phonological theory.. In John A Goldsmith, Jason Riggle, & Alan C. L. Yu (eds.), *The Handbook of Phonological Theory*, 401–434. John Wiley & Sons.

Escudero, Paola & Paul Boersma. 2001. Attested correlations between the perceptual development of L2 phonological contrasts and target-language production confirm the Gradual Learning Algorithm.. Philadelphia.

Haegeman, Liliane. 1987. Register Variation in English: Some Theoretical Observations. *Journal of English Linguistics* 20(2). 230–248. doi:10.1177/007542428702000207 (9 April, 2013).

Hayes, Bruce & Margaret MacEachern. 1998. Quatrain form in English folk verse. *Language* 64. 473–507.

Hayes, Bruce, Bruce Tesar & Kie Zuraw. 2003. OTSoft 2.1. http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Jarosz, Gaja. 2010. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of child language* 37(3). 565–606. doi:10.1017/S0305000910000103.

Niyogi, Partha. 2009. *The Computational Nature of Language Learning and Evolution.*. The MIT Press.