Introduction to constraint-based phonology

ACTL Summer School 2018
Adam J. Chong
a.chong@qmul.ac.uk

DAY 5
Variation II & Wrap up

What we've seen so far

- Classical OT
  - Constraint, ranking, categorical outcomes

We've seen two ways that a classical OT type of model can be modified, while maintaining 'strict-ranking', to account for variable outputs

- Partial-ranking
  - Constraint, rankings – but some are only partially ranked – each time you need to decide on a ranking, variable outcomes
- Stochastic OT (partially introduced)

Phonological variation

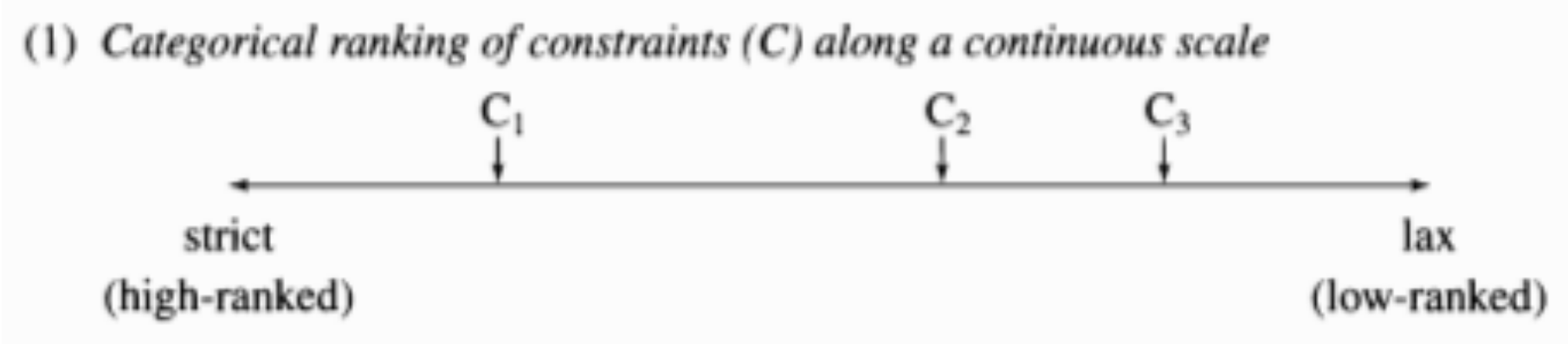
- The claim here is that (at least some) variation is **grammatical** – i.e. the degree of variability (so, say, relative proportion of forms) should be modeled in the grammar

- It can't just be boiled down to 'performance' (vs. 'competence') – you can't just put this into the bin of 'performance'

Stochastic OT (Boersma & Hayes, 2001)

- Still within the realm of 'strict'-ranking

- Instead of discrete rankings, constraints are on a **linear** scale of strictness



Stochastic OT (Boersma & Hayes, 2001)

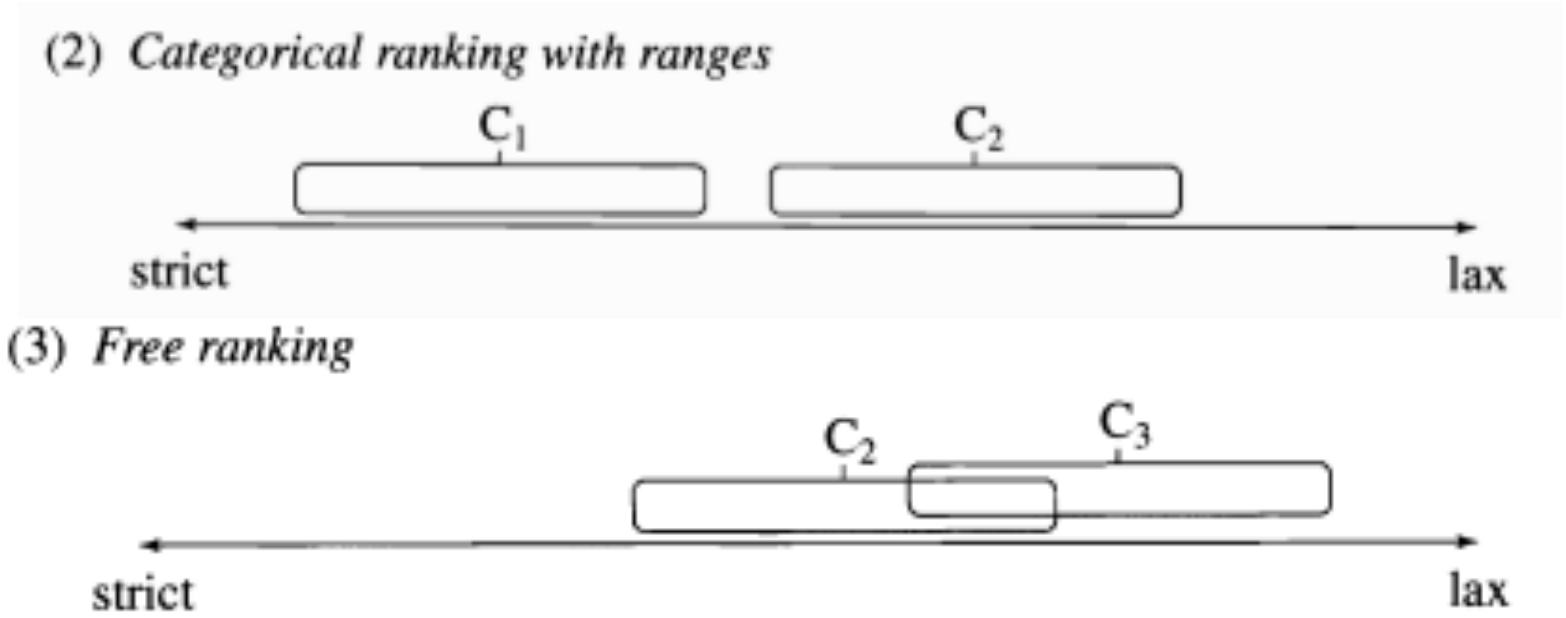
- Constraints are associated with a range of values instead of a fixed value

- When a speaker has to produce a form (i.e. at *evaluation time*), the position of constraints are perturbed by a random positive/negative value of noise (within the range of possible values)

- Choose one point from each constraint's range, then use a total ranking according to those points.

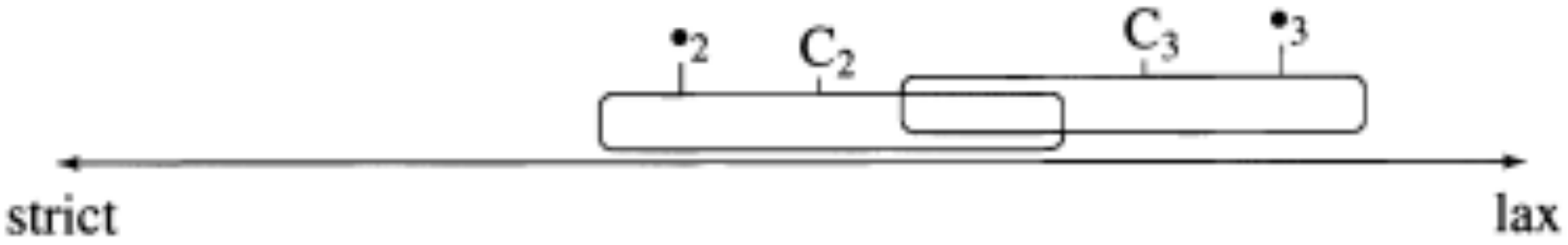
Stochastic OT (Boersma & Hayes, 2001)

- When a speaker has to produce a form (i.e. at *evaluation time*), the position of constraints are perturbed by a random positive/negative value of noise.
- Constraints are associated with a range of values instead of a fixed value

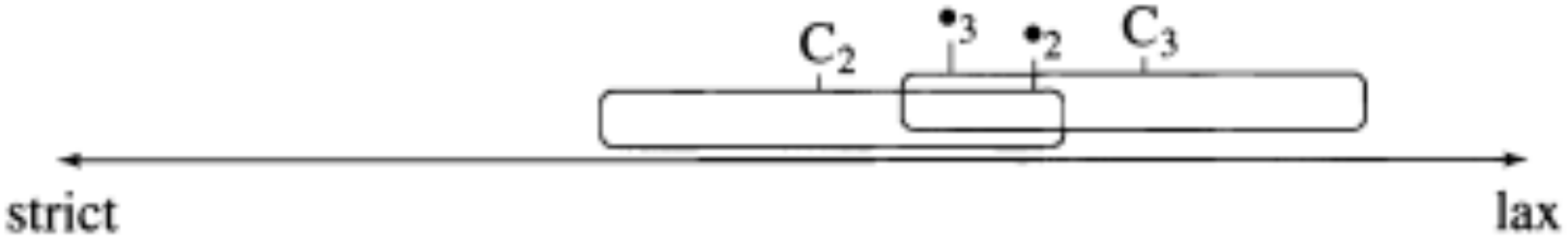


Stochastic OT (Boersma & Hayes, 2001)

(4) a. *Common result:* $C_2 \gg C_3$



b. *Rare result:* $C_3 \gg C_2$



- [demo in Excel]
- (from Kie Zuraw)

Example: Hungarian Vowel Harmony
(Hayes & Londe, 2006)

- Dative suffix [nɛk]~[nɔk]
- Basic pattern: allomorph selection is based on final vowel in stem.

(2)	BB	[ɔbl̲ɔk-nɔk]	*ablaknak*	'window-DAT'
	NB	[bi:ro:-nɔk]	*bírónak*	'judge-DAT'
	FB	[glyko:z-nɔk]	*glükóznak*	'glucose-DAT'

(3)	F	[yʃt-nɛk]	*üstnek*	'cauldron-DA
	BF	[ʃofø:r-nɛk]	*sofőrnek*	'chauffeur-D

Variability with vowel harmony

- Hayes & Londe focus on particular stem-type: cases where a stem-final neutral vowel is preceded by back-vowels
- These can either take the back or front suffixes, some also vacillate between both options.

(7)	BN	[pɔlleːr-nɔk]	*pallérnak*	'foreman-DAT'
	BN	[ɔrzeːn-nɔk, ɔrzeːn-nɛk]	*arzénnak, arzénnek*	'arsenic-DAT'
	BBN	[mutɔgeːn-nɛk]	*mutagénnek*	'mutagen-DAT'

Statistical generalizations

- **Height** effect: the lower the rightmost vowel, the more likely the front suffix is selected. E.g. %back is low with [ɛ] but highest with [i]

- **Count** effect: BN stems take back suffixes more than BNN stems (i.e. the further away the trigger for harmony is (B), the less likely you get harmony occurring

Statistical generalizations

tem ype	back	vacillator	front	total stems	backness index
	6251	39	0	6290	0·999
N	603	78	83	764	0·831
Bi	458	17	0	475	0·989
Bi:	52	0	1	53	0·980
Be:	93	18	9	120	0·845
Bɛ	0	43	73	116	0·104
NN	6	21	44	71	0·206
BNi	1	12	17	30	0·223
BNi:	1	7	0	8	0·358
BNe:	4	2	6	12	0·421
BNɛ	0	0	21	21	0
	14	23	259	296	0·078
N	0	4	939	933	0·002
	0	0	698	698	0

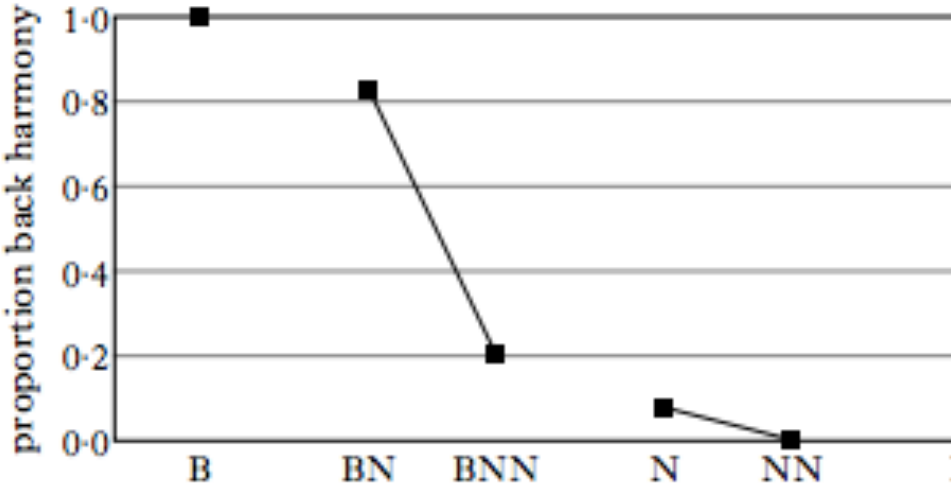


Figure 1

Google data: basic stem types.

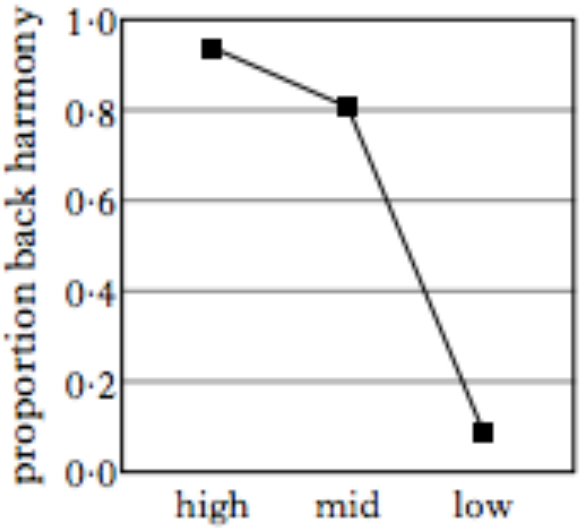


Figure 2

Google data: height effect.

Native speaker confirmation

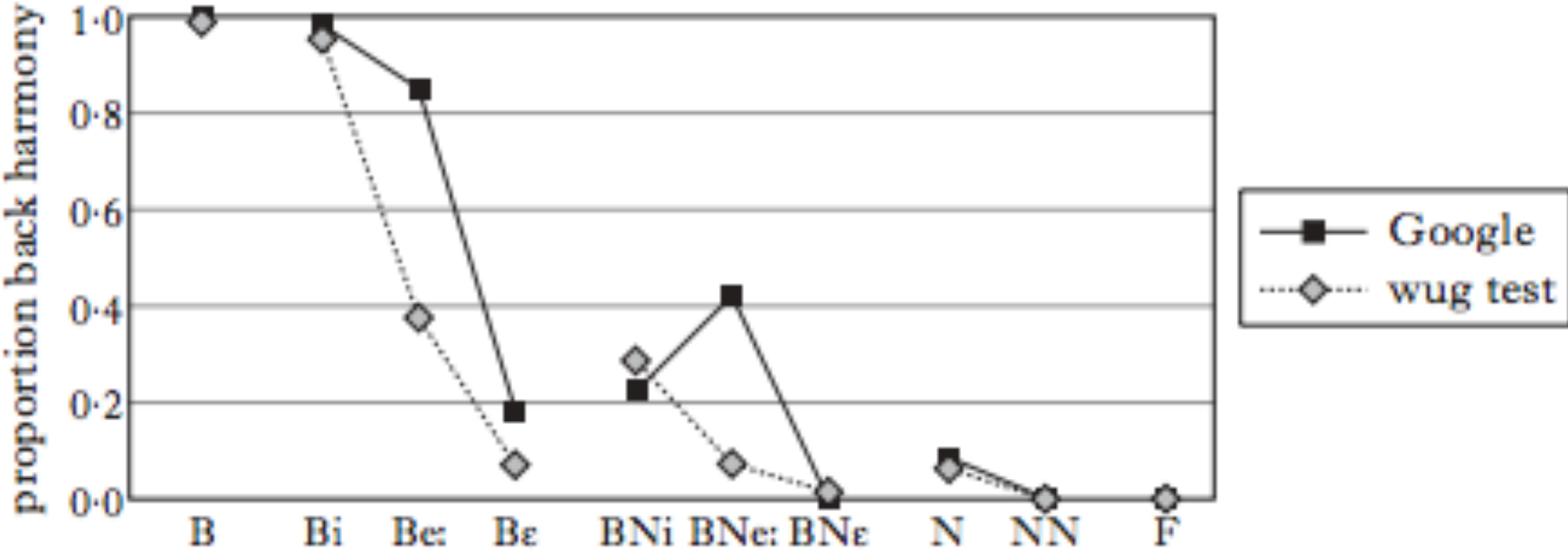


Figure 5

Wug-test data compared with Google data.

Final grammar

(without going through the full analysis)

- Using Stochastic OT: able to generate a grammar that generates outputs in proportions similar to those produced by native speakers!

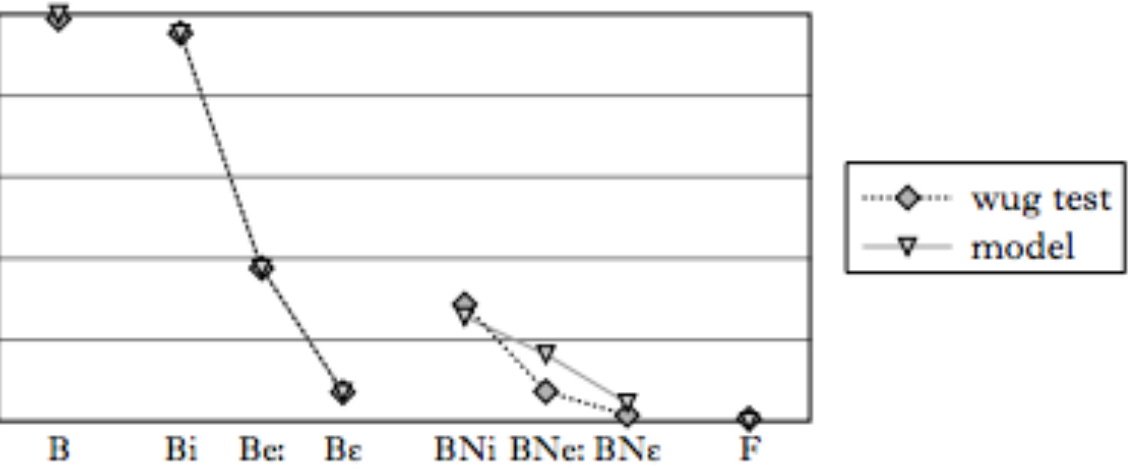
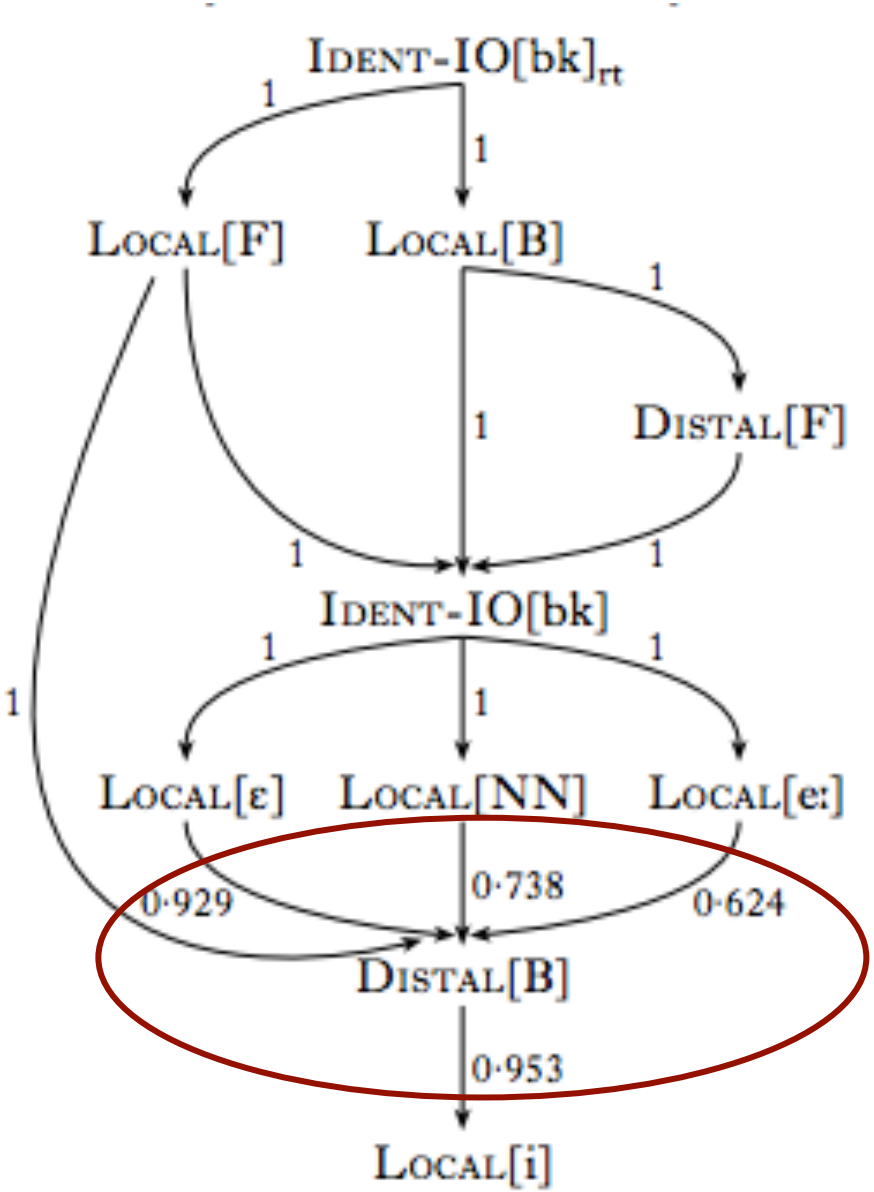


Figure 6
Match-up of model to wug-test data.

(34)



A few things to note

- Good example of triangulation: Hayes & Londe use data from:
  - Google search – lexical search
  - Confirmed with Wug test (test of productivity)
  - Tested model against data from Wug test
- Here, the model more-or-less replicates the behaviour of people!
- There's a learning algorithm for this – Gradual Learning Algorithm (Boersma, 1997) (though doesn't reliably converge – see Pater 2008)
- B&H 2001 also reanalyze the Anttila 1997 data using Stochastic OT with similar results

How is variation accounted for in constraint-based grammars?

- Brief look at four different models:
  - Partial-ranking
  - Stochastic OT

  } Use strict-ranking of constraints
  - Noisy Harmonic Grammar
  - MaxEnt Harmonic Grammar

What we've seen so far

- Rest of today: a small paradigm shift – using weights instead of strict domination

Harmonic Grammar

Actually a predecessor to OT: Legendre, Miyata & Smolensky (1990)

- Similar to OT in that it uses constraints
- But different: constraints are given **weights**

- This isn't a trivial difference – OT and HG predict different kinds of possible languages: specifically in relation to "ganging" and "cumulativity" effects.

Illustration: Coda devoicing in HG

Weight	1.5	1	$\mathcal{H}$
Input: /vrag/	*VOICED-CODA	IDENT-VOICE	
a. [vrag]	-1		-1.5
☞ b. [vrak]		-1	-1

- $\mathcal{H}$ = A candidate's Harmony score (the closer to positive, i.e. the higher, the more optimal)
- $\mathcal{H}$ = weight sum of constraint violations
- For [vrag] = (-1*1.5)+(0*1) = -1.5;
  For [vrak] = (0*1.5)+(-1*1) = -1

One key difference: HG predicts constraint 'ganging' effects that OT does not.

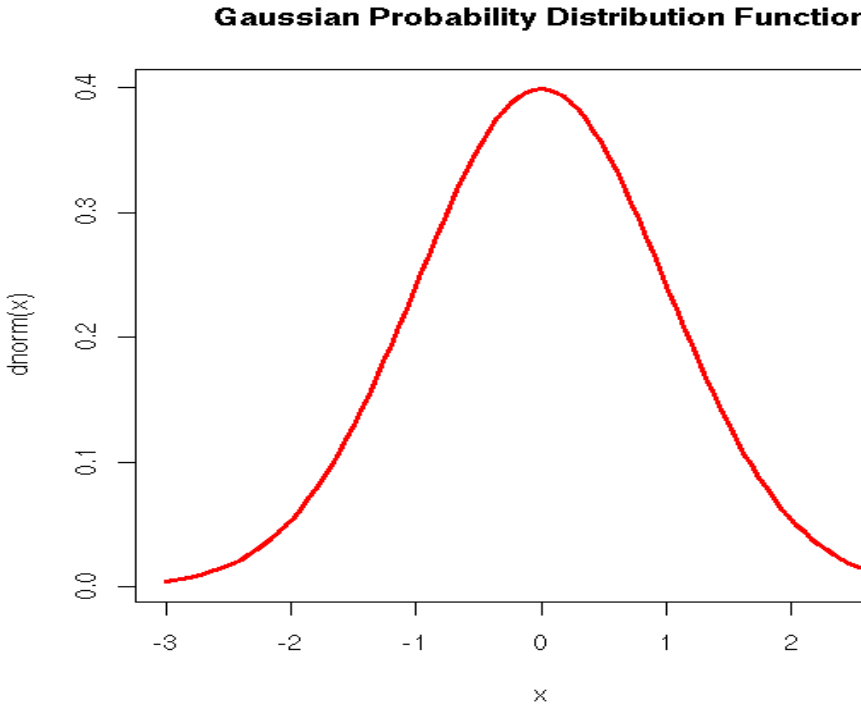
Illustrating 'ganging effects'

eight	3	2	2	$\mathcal{H}$
put: /input$_1$/	C_1	C_2	C_3	
a. Loser1	-1			-3
☞ b. Winner$_1$		-1		-2

eight	3	2	2	$\mathcal{H}$
put: /input$_1$/	C_1	C_2	C_3	
☞ a. Winner$_1$	-1			-3
b. Loser1		-1	-1	-4

How do you get variation here?

- Similar to Stochastic OT
- Every time evaluation of candidates occurs, the weight of each constraint is perturbed by a normally distributed (Gaussian) random positive/negative number
- So like in OT, if the **weights** of two constraints are close to each other, the relative weightings of two constraints can switch, predicting different outputs



zoonek2.free.fr/UNIX/48_R/07.html

(Optional) Coda devoicing in HG

- $\mathcal{H}$ (HARMONY) =
  For each constraint, sum of (weight+noise)*(no. of violations)

Weight	1.5 (-.4)	1 (.2)	$\mathcal{H}$
Input: /vrag/	*VOICED-CODA	IDENT-VOICE	
☞ a. [vrag]	-1		-1.1
b. [vrak]		-1	-0.9

Noise values in ( )

Weight	1.5 (.4)	1 (-.2)	$\mathcal{H}$
Input: /vrag/	*VOICED-CODA	IDENT-VOICE	
a. [vrag]	-1		-1.9
☞ b. [vrak]		-1	-1.2

Coetzee (2009): t/d-deletion

Table 2. Percent deletion in different contexts[5]

Relative deletion rate		Pre-V *west end*	Pre-Pause *west*	Pre-C *west side*
Pre-C > Pre-Pause > Pre-V	AAVE (Washington, DC)	29	73	76
	Jamaican English	63	71	85
	New York City English	66	83	100
	Tejano English	25	46	62
	Trinidadian English	21	31	81
Pre-C > Pre-V > Pre-Pause	Philadelphia English	38	12	100
	Chicano English	45	37	62
	Columbus English[6]	39	25	49

The constraints

(3) *CT — Assign one violation for every word that ends on the sequence […Ct] or […Cd].[9]

MAX — Assign one violation mark for each segment in the input that does not have a correspondent in the output (no deletion)

MAX-PRE-V — Assign one violation mark for each segment that appear in pre-vocalic context in the input, and that does not have a correspondent in the output (no deletion before a vowel)

MAX-PRE-PAUSE — Assign one violation mark for each segment that appear in pre-pausal context in the input, and that does not have a correspondent in the output (no deletion before a pause)

(4) a. Tejano English: Faithfull output in pre-consonantal context

'west bank'	100.4 (-.1) *CT	99.6 (.8) MAX	3.0 (-.1) MAX-PRE-V	0.8 (.9) MAX-PRE-PAUSE	H
☞ west bank	-1				-100.3
wes bank		-1			-100.4

b. Tejano English: Deletion output in pre-consonantal context

'west bank'	100.4 (-.2) *CT	99.6 (.2) MAX	3.0 (-.3) MAX-PRE-V	0.8 (-.1) MAX-PRE-PAUSE	H
west bank	-1				-100.2
☞ wes bank		-1			-99.8

- Excel demo: Noisy harmonic grammar
- OTSoft (introduced on Wednesday) now also has a learner for NHG weights – have a look at home!
- Note that Coetzee (2009) uses Praat's implementation
  - Yes – you can do constraint grammars in Praat!

Table 3. Results of learning simulations

	Constraint weights					Deletion rates		
	*CT	MAX-PRE-V	MAX-PRE-PAUSE	MAX		Pre-V	Pre-Pause	Pre-C
AAVE	101.0	3.9	-1.7[12]	99.0	Obs	29	73	76
					Exp	28	72	75
Jamaica	101.4	1.7	0.7	98.6	Obs	63	71	85
					Exp	63	71	85
NYC	141.1	80.9	78.9	58.9	Obs	66	83	100
					Exp	65	83	100
Tejano	100.4	3.0	0.8	99.6	Obs	25	46	62
					Exp	26	45	61
Trinidad	101.2	5.2	4.2	98.8	Obs	21	31	81
					Exp	20	29	79
Philadelphia	139.2	79.5	82.4	60.9	Obs	38	12	100
					Exp	38	12	100
Chicano	100.4	0.9	1.8	99.6	Obs	45	37	62
					Exp	45	38	62
Columbus	100.0	0.1	2.1	100.1	Obs	39	25	49
					Exp	38	24	49

Incorporating ‘social’ factors

- Higher frequency words more likely that lower frequency words to show deletion (quite a common observation!)
- How should we model this?

- Coetzee’s answer: frequency effect as scaling factors (additional adjustments to constraints) specifically on faithfulness constraints

- Intuition: higher frequency words don’t have to be as “faithful”

Adding another factor into the equation – scaling factor!

(8) a. Tejano English: High frequency *just*

'just'	*wt*	*nz*	*wt*	*nz*	*sf*	*wt*	*nz*	*sf*	*wt*	*nz*	*sf*	
	100.4	0.2	99.6	0.9	-1.0	3.0	-0.3	-1.0	0.8	-0.8	-1.0	
	*Ct		Max			Max-Pre-V			Max-Pre-Pause			H
just	-1											-100.6
☞ jus			-1						-1[14]			-99.5

b. Tejano English: Low frequency *jest*

'jest'	*wt*	*nz*	*wt*	*nz*	*sf*	*wt*	*nz*	*sf*	*wt*	*nz*	*sf*	
	100.4	0.2	99.6	0.9	1.0	3.0	-0.3	1.0	0.8	-0.8	1.0	
	*Ct		Max			Max-Pre-V			Max-Pre-Pause			H
☞ jest	-1											-100.6
jes			-1						-1			-102.5

Summary so far

- We've seen how a basic OT grammar can be used/augmented to account for variable outputs
- Allows for modeling of quantitative patterns – experimental/corpus data
- Some software that can be used to look at this:
  - OTSoft (Windows)
  - OTHelp (cross-platform)
  - OTWorkplace (Windows)
  - MaxentGrammar Tool (cross-platform)

Maximum Entropy Harmonic Grammar

- Goldwater & Johnson, 2003: applied widely-used machine learning technique to constraint grammars (see also Hayes & Wilson, 2008)
- Similar to regular HG (weighted constraints)
- But whereas in noisy HG, noise has to be added to the weights of the constraints to enable us to capture variation
- In MaxEnt HG, the weighted sum (harmony score) is exponentiated: e^weighted_sum
  - From this we can then get the **probability** of each candidate (proportional to that number)
- What's also nice about MaxEnt HG is that it is mathematically really well understood – basically logistic regression!

Connecting this to grammar: Maximum Entropy grammars

Just like Harmonic Grammar, except:

- in HG, harmony is the weighted sum of constraint violations
  - Candidate with best harmony wins
  - We need to add noise to weights in order to get variation
- In MaxEnt, we *exponentiate* the weighted sum: $e^{weighted\_sum}$
  - Each candidate's probability of winning is proportional to to the sum of this score across all the candidates

$$Pr(y|x) = \frac{exp(\sum_{i=1}^{m} w_i C_i(y,x))}{Z}, where$$

$$Z = \sum_{y \in Y(x)} \frac{exp(\sum_{i=1}^{m} w_i C_i(y,x))}{Z}$$

Noisy HG reminder

/aktmo/	*CCC w= 5	Max-C w = 4	Dep-V w = 3	NoisyHG harmony
[aktmo]	*			-5
[akitmo]			*	-3
[aktimo]			*	-3
[akmo]		*		-4

Maximum Entropy

aktmo/	*CCC w= 5	Max-C w = 4	Dep-V w = 3	MaxEnt score	MaxEnt prob. of winning = harmony/sum
aktmo]	*			e^{-5}	0.05
akitmo]			*	e^{-3}	0.40
aktimo]			*	e^{-3}	0.40
akmo]		*		e^{-4}	0.15
				sum = 0.125 (Z)	

- Let’s see the NHG tableaux with Coetzee’s data but now in MaxEnt

Differences from Noisy HG

- In MaxEnt, it's a bit easier to see each candidate's probability—we can calculate it directly
- There really isn't a 'noise' factor as such
- The mathematics is better understood/connected = logistic regression

- Difficult to find cases where the results are clearly distinguished (see Hayes, 2017)

Software

- OTSoft doesn't implement MaxEnt HG

- You'll have to use MaxEnt Grammar Tool (Java implemented, so cross-platform – not just Windows!)
  - Similar input files
  - Also available online: http://linguistics.ucla.edu/people/hayes/MaxentGrammarTool/

For those interested in interaction of socio-factors in such a model

- Franny Brogan's very very very recent dissertation from UCLA (Hispanic Linguistics) - 2018
- Looking at /s/-lenition in Salvadoran Spanish
- Coetzee-esque use of scaling factors but using MaxEnt instead of NHG

Other software

- OT Workplace: https://sites.google.com/site/otworkplace/
- OT Help: https://people.umass.edu/othelp/

Important note

- Just because there is software that can calculate weights etc, this doesn't mean the phonologist is out of a job!

- You still need to come up with the relevant constraints!!

- The question of how constraints are learnt – not a trivial issue.
  - E.g. Hayes & Wilson (2008)

- But – given a set of constraints, both MaxEnt and NHG have reliable learning algorithms that will learn the set of weights that fit the input data

Other things to note

- What's been lurking here is also the question of learning
- How are constraint rankings/weights learnt?
- How well do the learnt weights/rankings fit with the observed data
- MaxEnt, NHG, and Stochastic OT have learning algorithms

Summary for the week

Day 1-3:

- Gave a brief history and motivation for why constraint-based frameworks were adopted by many phonologists
- In particular: Optimality Theory
  - Constraint interaction – what does this buy us?
  - TETU, contrasts, conspiracies

Day 4-5:

- Constraint-based grammars – how can these be used to deal with phonological variation?

Where to go: (not exhaustive!)

- All the references from Yuni's class
- More theoretical phonology – various patterns/issues
- How to deal with variation – grammatical/non-grammatical?
- Phonological learning – how are constraints learnt? Constraint rankings? How can a constraint-grammar be used to understand what is going on in learning?
- Speech perception – e.g. work by Paul Boersma and colleagues on Bi-directional OT
- Serial versions of OT – Harmonic Serialism
  - Where does OT fall short? (Opacity)
- Representations?
- Experimental evidence? How to incorporate insights into models?

Thanks for attending!

- Stay in touch: a.chong@qmul.ac.uk

- Working on a phonology/phonetics project (with or without variation)? I'd be really interested to discuss it with you/read it (esp. if there's an experiment involved, or it's on an understudied language).

- If you're in the London area, don't hesitate to make an appointment to meet on QMUL campus (Mile End)!

Some references

- Legendre, Miyata, & Smolensky 1990: original proposal
- Smolensky & Legendre 2006: a book-length treatment
- Pater & Boersma 2008: noisy HG (for non-varying data)
- Pater, Jesney & Tessier 2007; Coetzee & Pater 2007: noisy HG for variation
- Coetzee, A. (2009). An integrated grammatical/non-grammatical model of phonological variation. In Kang et al. (eds.). *Current issues in Linguistic interfaces.* See also Coetzee (2016) in *Phonology*
- Goldwater & Johnson (2003), Hayes, B. & Wilson, C. (2008) on MaxEnt