CS 585: INTRODUCTION TO NATURAL LANGUAGE PROCESSING

Guest Lecture: Gaja Jarosz

Gradient Phonotactics

- How good are these as novel words of English?
  - stin
  - blick
  - mip
  - skell
  - blafe
  - bwip
  - shmy
  - smum
  - dlap
  - bzack
  - mrock
  - dmell
  - lnoot
  - mdap

Gradient Phonotactics

- How good are these as novel words of English?
  - stin
  - blick
  - mip
  - skell
  - blafe
  - bwip
  - shmy
  - smum
  - dlap
  - bzack
  - mrock
  - dmell
  - lnoot
  - mdap

Gradient Phonotactics

- Knowledge of gradient phonotactics has been illustrated in a range of tasks
  - Lexical access
    - Faster recognition of more phonotactically probable words
  - Word learning
    - Faster acquisition of more phonotactically probable words
  - Word segmentation by children and adults
    - This is the main information we think infants use to segment speech
    - Example: Mattys & Jusczyk (2001)
      - ..bean gaffe hold.. [ng] and [fh] occur infrequently within words
      - ..fang gaffe tine.. [ŋg] and [ft] occur frequently within words

What's Phonotactic Probability?

- What's the probability of?
  - blick vs. shmy vs. mrock
- Ideas?

N-grams!

- Pr(#abcd#)
- Pr(a | b) = ?
- Chain Rule
  - Pr(a,b) = Pr(a | b)Pr(b)
  - Pr(a,b,c) = Pr(a | b,c)Pr(b | c)Pr(c)
    = Pr(c | a,b)Pr(b | a)Pr(a)
- Apply chain rule to Pr(#abcd#)?

N-grams!

- Pr(#abcd#)?
  - Pr(a | #) *
  - Pr(b | #,a) *
  - Pr(c | #,a,b) *
  - Pr(d | #,a,b,c) *
  - Pr(# | #,a,b,c,d)
- N-grams make an independence assumption that only a fixed amount of history matters…
  - Unigrams, bigrams, trigrams, etc.

N-grams

- We can't condition on every possible preceding sequence - there's too many!
  - Jack and Jill went up the hill
- N-gram models condition on N-1 previous symbols:
  - Unigrams: no history
    - P(Jack)P(and)P(Jill)P(went)P(up)P(the)P(hill)
  - Bigrams: 1 previous symbol
    - P(Jack | #)P(and | Jack)P(Jill | and)P(went | Jill)…
  - Trigrams: 2 previous symbols
    - P(Jack | ##)P(and | # Jack)P(Jill | Jack and)P(went | and Jill)…

What's Phonotactic Probability?

- What's the bigram probability of?
  - blick vs. shmy vs. mrock

What's Phonotactic Probability?

- What's the bigram probability of?
  - blick vs. shmy vs. mrock
- P(b | #)P(l | b)P(i | l)…
  - How can we estimate these probabilities?

Baseline: Phoneme bi-grams

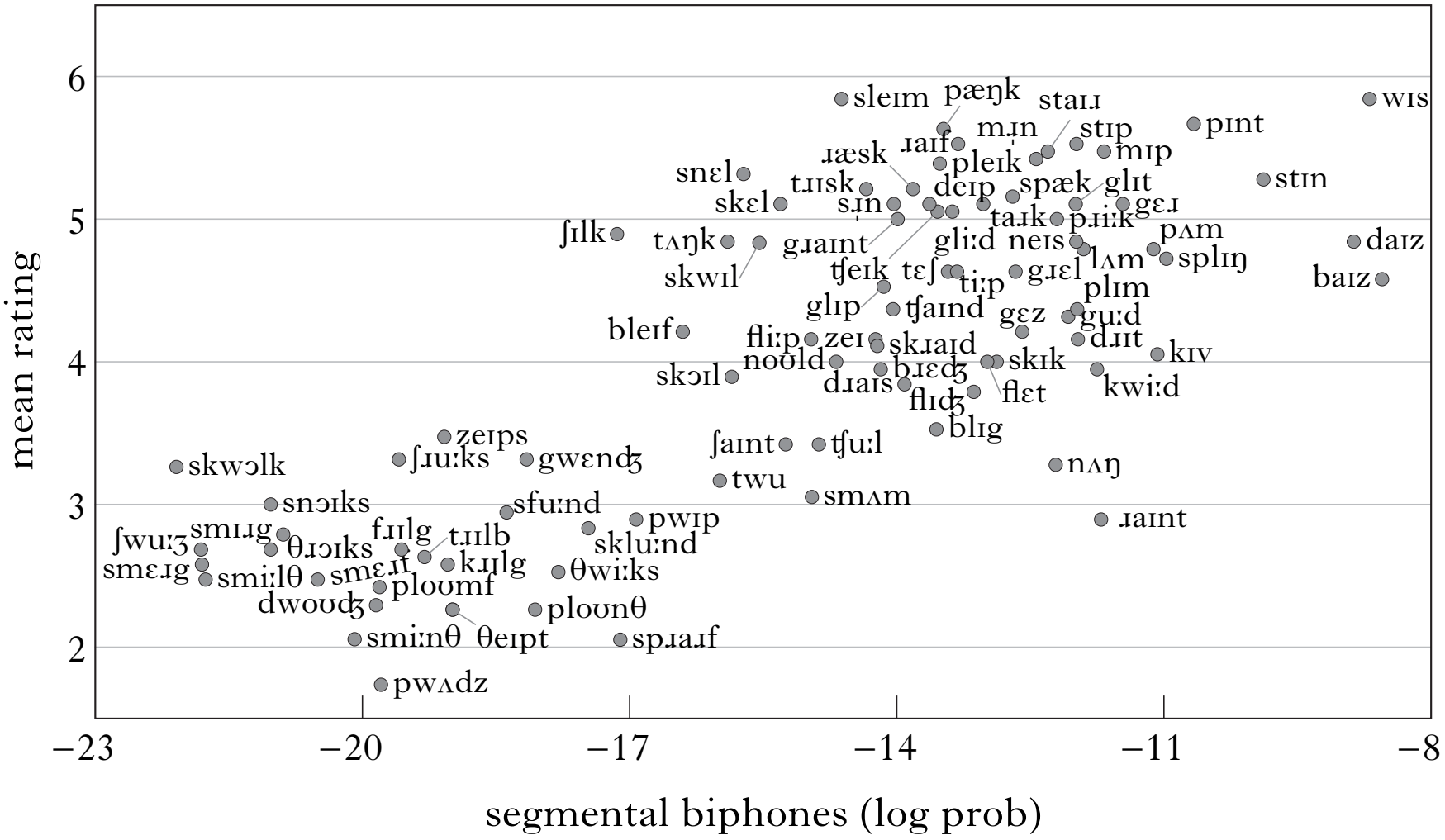
- Example

biphone	stɪn		bleɪf	
1	#s	0·118	#b	0·057
2	st	0·205	bl	0·106
3	tɪ	0·192	leɪ	0·042
4	ɪn	0·108	eɪf	0·007
5	n#	0·151	f#	0·067
log (prob)		–4·12		–6·93

What's Phonotactic Probability?

- What's the bigram probability of?
  - blick vs. shmy vs. mrock
- If we're primarily interested in the initial consonant sequences, there are several ways we could use n-grams to model these sequences…
  - Pros and cons?

How do Bigrams do? (Albright 2009)



How do bigrams do?

- Overall correlations with human judgments?
  - Training on just onsets:
    - r = 0.88
      - Hayes & Wilson (2008)
  - Training on whole words:
    - r = 0.78 (with syllable boundaries), = .50 (without)
      - Daland et al. (2011)

Problem!

- People show gradient judgments of two kinds
- Attested onset clusters
  - stin > blin > bwin
- And unattested onset clusters
  - bnick > bzick > mbick
- High correlations due to differences across not within
- This latter category gives us crucial clues about how people *learn and represent* these patterns
- However,
  - Positional bigrams cannot model this. Why?
  - Whole-word bigrams do very poorly (r = 0.22). Why?

So what are people doing?

- Why do people have these judgments then?
  - bnick > bzick > mbick
- Notice: this is a problem with generalization
  - We want models to generalize correctly to new data
  - Cognitive scientists & Linguists are interested in modeling how humans generalize (correct = whatever humans do)
  - N-grams already generalize, just not in the right way!
- Ideas? Intuitions? Can we save this approach?
  - Hint: this is not a solved problem yet!

So what are people doing?

- Why do people have these judgments then?
  - bnick > bzick > mbick
- Two recent models
  - Albright 2009 featural bigrams
  - Hayes & Wilson 2008 Maximum Entropy model

Albright's Proposal

- bn is relatively good because it's similar to existing
  - bl
  - br
  - sn
  - …
- bd is not as similar to existing sequences
  - bl is farther from bd than from bn
  - br is farther from bd than from bn
  - …
- How to formalize *similar*?
  - Phonetic/phonological features!

Phonological Features

		diacritic examples			Coronal Obstruents [+cons, -son, +cor]										Palatal Obstruents [+cor + dors]				Non-coronal Obstruents [+cons, -son, -cor]																Laryngeals [-cons,-son]		
		t’	t^h	t̪	t	d	s	z	ɬ	ɮ	θ	ð	ʃ	ʒ	c	ɟ	ç	ɟ̑	p	b	f	v	ɸ	β	k	g	x	ɣ	q	ɢ	χ	ʁ	ħ	ʕ	h	ɦ	ʔ
Class features	**cons**	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-
	son	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	syll	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Place features	**labial**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
	round	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0
	coronal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	ant	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	distrib	-	-	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	dorsal	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	-
	high	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	0	0	0	0	0	0	+	+	+	+	-	-	-	-	0	0	0	0	0
	low	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0	0	0	0	0
	back	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0	0	0	0	0
	tense	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	0	0	0	0	0	0	-	-	-	-	-	-	-	-	0	0	0	0	0
	phrngl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
	ATR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	-	0	0	0
Laryngeal features	**voice**	-	-	-	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	-	-	+	-
	S.G.	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-
	C.G.	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
Manner features	**cont**	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	+	+	-	-	+	+	+	+	-	-	+	+	-	-	+	+	+	+	+	+	-
	strident	-	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	lateral	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	del rel	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Albright's Proposal

- Do bigrams over phonological *classes* not phonemes
- bn is b[+voi, +son, +cor, -strid, +nas]
  - bl is b[+voi, +son, +cor, -strid, ~~+nas~~]
    - We can group n and l into a small natural class b[n l r j]
    - Frequency of this specific combination is high
      - bl, br, bj are all attested
- bd is b[+voi, -son, +cor, -cont]
  - bl is b[+voi, ~~-son,~~ +cor, ~~-cont~~]
    - l and d have less in common, need larger class b[n l r j z n d]
    - Frequency of this combination is high, but not very related to bd

Albright's Proposal: Formally

- Phonotactic probability tradeoff between frequency and specificity
  - $score(ab) = \max_{A,B \in Nat} \frac{count(AB)}{count(..)} \times P(a \mid A) \times P(b \mid B)$
  - Where
    - $P(a \mid A) = \frac{1}{|A|}$
  - How does this deal with bd vs. bn vs. bl?

Hayes & Wilson (2008)

- Maximum Entropy Model of Phonotactics
  - closely related to regression models (in fact they are equivalent to multinomial logistic regression models)
  - You have some observation c to assign probability to
  - $$P(y = c) = \frac{e^{f_c(x)}}{Z} = \frac{e^{(w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n)}}{Z}$$
  - You do so on the basis of various features $x_1 \ldots x_n$
- For phonotactics, these features reflect phonological preferences and restrictions.
  - In Linguistics, features are called constraints
  - Constraint-based models of various kinds are widely used in theoretical phonology

Top MaxEnt Constraints (HW 2008)

	Constraint	Weight	Comment	Examples
1.	*[+son, +dors]	5.64	*[ŋ]	*ŋ, *sŋ
2.	*[+cont, +voice, −ant]	3.28	*[ʒ]	*ʒ (see also #16)
3.	$*\begin{bmatrix}\wedge -\text{voice}\\ +\text{ant}\\ +\text{strid}\end{bmatrix}[-\text{approx}]$	5.91	Nasals and obstruents may only be preceded (within the onset) by [s].	*kt, *kk, *skt
4.	*[ ][+cont]	5.17	Fricatives may not cluster with preceding C.	*sf, *sθ, *sh, *sfl
5.	*[ ][+voice]	5.37	Voiced obstruents may not cluster with preceding C.	*sb, *sd, *sgr
6.	*[+son][ ]	6.66	Sonorants may only be onset-final.	*rt
7.	*[−strid][+cons]	4.40	Nonstrident coronals may not precede nonglides.	*dl, *tl, *θ1
8.	*[ ][+strid]	1.31	Stridents must be initial in a cluster.	*st͡ʃ (see also #14, #22)
9.	$*[+\text{lab}]\begin{bmatrix}\wedge +\text{approx}\\ +\text{cor}\end{bmatrix}$	4.96	The only consonants that may follow labials are [l] and [r].	*pw vs. pl, pr
10.	$*[-\text{ant}]\begin{bmatrix}\wedge +\text{approx}\\ -\text{ant}\end{bmatrix}$	4.84	Only [r] may follow nonanterior coronals.	*ʃl vs. ʃr
11.	*[+cont, +voice][ ]	4.84	Voiced fricatives must be final in an onset.	*vr, *vl vs. fr, fl
12.	*[−cont, −ant][ ]	3.17	[t͡ʃ] and [d͡ʒ] must be final in an onset.	*t͡ʃr, *d͡ʒr vs. tr, dr (see also #22)

Results Comparison (Daland et al 2011)

	syllabification				no syllabification			
model	at-tested	mar-ginal	unat-tested	over-all	at-tested	mar-ginal	unat-tested	over-all
Albright	**0·21**	**0·03**	**0·55**	**0·51**	0·13	−0·07	0·18	0·26
bigram	**0·19**	**0·16**	**0·22**	**0·78**	0·23	0·01	−0·14	0·50
Coleman	**0·35**	**0·31**	**−0·01**	**0·55**	–	–	–	–
gnm.fig	0·07	0·25	−0·29	0·15	0·06	0·24	−0·32	0·08
gnm.oral	0·28	0·23	−0·28	0·22	0·26	0·23	−0·28	0·21
gnm.writ	0·17	0·24	−0·17	0·24	0·16	0·24	−0·30	0·15
gnm.lin	**0·32**	**0·23**	**−0·22**	**0·31**	0·30	0·22	−0·26	0·24
hw[100]	**0·00**	**0·02**	**0·76**	**0·83**	0·00	−0·31	0·79	0·68
hw[150]	0·00	0·06	0·69	0·82	0·00	0·04	0·67	0·75
hw[200]	−0·09	0·03	0·64	0·80	0·00	0·05	0·69	0·77
hw[250]	−0·09	0·13	0·64	0·84	0·00	0·00	0·70	0·80
hw[300]	−0·39	0·04	0·54	0·80	0·00	−0·02	0·70	0·81
hw[350]	−0·39	0·03	0·51	0·80	0·00	−0·10	0·67	0·81
hw[400]	−0·39	0·04	0·52	0·81	0·00	0·00	0·68	0·80
vl.uni	0·27	0·11	0·38	0·43	0·30	0·19	0·34	0·36
vl.bi	**0·30**	**0·06**	**0·27**	**0·56**	0·30	0·08	0·22	0·54

Results Discussion

- Observations
  - Bigram (.78) and Maxent (.83) do well overall
    - But it's easy to do well overall
  - Bigrams (.22) do terribly on unattesteds
  - MaxEnt does well on unattesteds (.76)
  - No model does great on attesteds
- Conclusions?
  - Long way to go…
  - But…
    - Features/similarity are important
    - Syllabification/higher level structure is important
  - Ideas? Critiques? Questions?

Other Work

- One example of computational modeling to address cognitive questions
  - Main argument: human phonological generalization is feature-based and structure sensitive
- Modeling can also be used to address questions about learning *bias* to show how humans systematically ignore or depart from certain properties of the input

Sonority Sequencing Principle

- Sonority Sequencing Principle (SSP)
  - [lb]ack ≺ [mb]ack ≺ [bd]ack ≺ [bn]ack ≺ [bɹ]ack ≺ [bj]ack
  - -2 -1 0 1 2 3
  - Typological Generalizations (Clements 1988, Selkirk 1984)
    - [bn]ack ⇒ [bl]ack
- Consistent findings of **Sonority Projection** in English
  - Preferences between unobserved clusters (nb vs db)
  - Production, perception, acceptability; aural, written

    (Berent et al. 2007, Berent & Lennertz 2009, Berent et al. 2009, Davidson et al. 2004, Davidson 2006, Daland et al. 2011)
- Debate: Where do these preferences come from?
  - Daland et al examine SSP and argue models like MaxEnt can capture these effects…

How can this be? Statistics over classes!

OO		ON		OL		OG	
st	521	sn	109	pr	1046	kw	201
sp	313	sm	82	tr	515	sw	153
sk	278			kr	387	hw	111
				gr	331	tw	55
				br	319	dw	17
				fl	290	gw	11
				kl	285	θw	4
				fr	254		
				pl	238		
				bl	213		
				sl	213		
				dr	211		
				gl	131		
				θr	73		
				ʃr	40		
1112		**191**		**4546**		**552**	
(17.4%)		**(3.0%)**		**(71.0%)**		**(8.6)%**	

- Plenty of SSP evidence
  - LO none
  - NO none
  - OO only sO
  - ON only sN
  - OL 71%
- Higher level
  - Son Rise 82.6%
  - Son Plat. 17.4%
  - Son Fall 0%

Polish Clusters

- My ongoing work
  - English is the wrong language to test!
    - Predictions based on input and based on bias are the same!
- Change tactic: look at a language that admits all these sequences!
  - What are predictions based on input?
  - What are predictions based on universal bias?
  - Build models for both and evaluate!
- Polish onsets
  - Whole Scale!
  - [wb]ack ≺ [lb]ack ≺ [mb]ack ≺ [bd]ack ≺ [bn]ack ≺ [bɹ]ack ≺ [bj]ack
  - -3 -2 -1 0 1 2 3

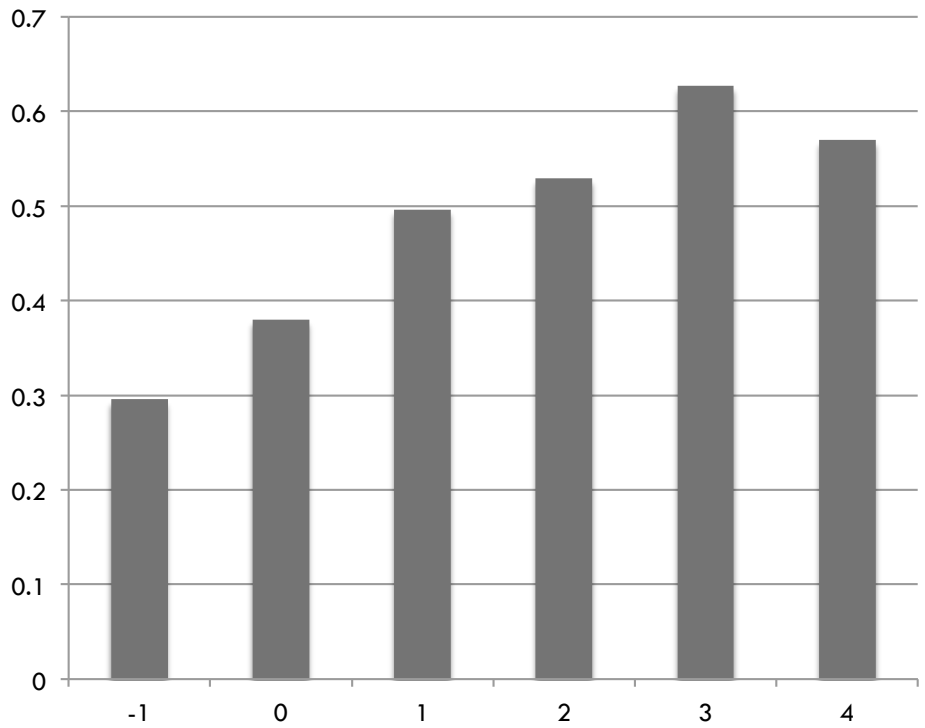
Polish Clusters: Examples

	P	F	N	L	G
P	[ptak] "bird"	[pʂɨjɕt͡ɕ] "to come"	[dnɔ] "bottom"	[kluʧ] "key"	[gwɔva] "head"
	[ktɔ] "who"	[kɕɔw̃ʂ̃ka] "book"	[dɲa] "day"	[drɔga] "road"	[bjawɨ] "white"
F	[ɕpi] "sleeps"	[xfila] "moment"	[ɕɲɛk] "snow"	[vlat͡ɕ] "to pour"	[vwaɕɲɛ] "exactly"
	[stɔ] "one hundred"	[fʂɨstkɔ] "everything"	[smɔk] "dragon"	[fruvat͡ɕ] "to fly"	[zwɨ] "bad"
N	[mdwɔ] "dull"	[mʂa] "mass"	[mɲɛ] "me (inst.)"	[mlɛkɔ] "milk"	[mjaw] "he had"
	[mgwa] "fog"	[mʂɨt͡sa] "mite"	[mnɔga] "multiple"	[mrufka] "ant"	[mjut] "honey"
L	[rtɛɲt͡ɕ] "mercury"	[lvɨ] "lions"	[lɲanɨ] "linen" (adj)		
	[rd͡za] “rust”	[rvat͡ɕ] “tear”	[lnu] "linen" (gen)		
G	[wba] "head (gen)"	[wza] "tear"			
	[wkat͡ɕ] "sob"				

What do kids learning Polish do?

SSP effects using scale: P≺F≺N≺L≺G≺V

This is how accurate children are at producing onset clusters as a function of sonority rise degree. Strong relationship!



Polish Clusters: statistics

- Could kids get these preferences from input statistics?
  - Relative frequency very different from English
  - Half the input is Obstruent-Obstruent!
- I consider various ways of modeling the input
  - Segment bigrams
  - Class-based bigrams
  - MaxEnt
- None can capture the kids' preferences…

Sonority Profile	LO	**OO**	NN	ON	NL	OL	NG	OG
Token Frequency	0.04%	**50.90%**	0.50%	3.60%	0.20%	20.40%	3.00%	21.30%
Type Frequency	0.10%	**47.70%**	0.20%	6.70%	0.60%	19.60%	2.90%	22.30%

Thanks!

- Two brief examples of what we do
- Lots of other people in Linguistics interested in computational modeling (not just phonology)
- If you're interested in more…
  - Come talk to us!
  - Check out our classes!