HIDDEN STRUCTURE AND AMBIGUITY IN PHONOLOGICAL LEARNING

GAJA JAROSZ

Society for Computation in Linguistics (SCiL) 2019

January 5, 2019, NYC

TWO STRANDS OF PROGRESS

Increasingly realistic assumptions about learning

- Hidden Structure & Ambiguity
- Quantitative Patterns & Generalizations

Quantitative modeling is an integral component of both

Both have led to methodological advancements

- Enhanced modeling capabilities
- Novel empirical connections
  - Richer learning data: Corpora
  - Richer assessment data: Behavioral Data
- Qualitative paradigm shift: gradience in learn(ing/ability)
  - Role of learning in phonological theory

OVERVIEW

Embracing Ambiguity & Uncertainty

Gradience in Learn(ing|ability)

New Connections & Resulting Discoveries

- (soft) Biases
- Explanatory role of learning

New questions and under-explored directions

EMBRACING AMBIGUITY & UNCERTAINTY

Inconsistency

- Noise & Errors
- Exceptions
- Quantitative Generalizations
  - Free variation, gradient phonotactics, patterned exceptionality

Hidden Structure

- Prosodic structure (feet, syllables, autosegmental structure…)
- Underlying representations
- Segmentation (morphemes, words)
- Derivational Ordering
- Rules & Constraints
- Exceptionality (Classes)
- …

EMBRACING AMBIGUITY & UNCERTAINTY

Ambiguity $\Rightarrow$ Uncertainty

Uncertainty $\Rightarrow$ Decisions

- What do learners do when there are multiple options?

Balancing and Integrating conflicting pressures

- Generalize or memorize?
- Where to attribute generalizations?
- Accumulating knowledge in the face of ambiguity

Understanding how learners do this, examine

- *Generalizing*
- *like humans* from
- *finite sample of imperfect, ambiguous, gappy* data

EMBRACING AMBIGUITY I: GRADIENT PHONOTACTICS

English Initial Clusters

st	521	sn	109	fl	290	pɹ	1046
sp	313	sm	82	kl	285	tɹ	515
sk	278			pl	238	kɹ	387
				bl	213	gɹ	331
				sl	213	bɹ	319
				gl	131	fɹ	254
						dɹ	211
						kw	201
						sw	153
						hw	111
						θɹ	73
						tw	55
						ʃɹ	40
						dw	17
						gw	11
						θw	4

How do speakers generalize phonotactics?

- One pressure: tightly fit the data. Learn restrictions!
- Conflicting pressure: generalize to unseen data!

Experimental findings: gradient generalization

- ‘mip’ > ‘bwip’ > ‘dlap’ > ‘bzap’
- Coleman & Pierrehumbert 1997, Bailey & Hahn 2001, Davidson 2007, Berent et al. 2007, Hayes & Wilson 2008, Albright 2009, Daland et al. 2011, …

(data from Hayes & Wilson 2008)

EMBRACING AMBIGUITY I: GRADIENT PHONOTACTICS

English Initial Clusters

st	521	sn	109	fl	290	pɹ	1046
sp	313	sm	82	kl	285	tɹ	515
sk	278			pl	238	kɹ	387
				bl	213	gɹ	331
				sl	213	bɹ	319
				gl	131	fɹ	254
						dɹ	211
						kw	201
						sw	153
						hw	111
						θɹ	73
						tw	55
						ʃɹ	40
						dw	17
						gw	11
						θw	4

Quantitative modeling

- Captures gradience
- Formalizes balance: fit and generalization
- Formalizes ‘similar enough’
  - Necessary even for categorical generalizations!

How is generalization constrained?

- What representations underlie generalization?
- What principles underlie generalization?

(data from Hayes & Wilson 2008)

A CONTINUUM OF GENERALIZATIONS

Default Hypothesis: lexical statistics – but how?

Increasingly Rich Hypotheses. Frequency ++…

- **Segmental statistics**, no similarity
  - Analogy (Bailey & Hahn 2001)
  - Phoneme co-occurrence (Vitevich & Luce 2004)
- **+ Class-Based Generalization (CBG)**
  - Abstract representations: features, syllables, tiers, etc.
  - UCLA Phonotactic Learner (Hayes & Wilson 2008)
  - Featural Bigram Model (Albright 2009)
- **+ Universal Bias**
  - Inherent preferences among abstract representations:
    - Sonority Plateau (#bd) $\prec$ Sonority Rise (#bl)

CORRELATIONS (JAROSZ & RYSLING 2016)

	Unsyllabified			Syllabified		
	Overall	*Attest*	*Unattested*	*Overall*	*Attest*	*Unattested*
Grapheme Bigram	**0.65**	**0.52**	0.20			
Grapheme Trigram	**0.84**	**0.84**	-0.03			
Phoneme Bigram	**0.63**	**0.37**	0.15	**0.79**	**0.47**	0.15
Phoneme Trigram	**0.78**	**0.69**	-0.21	**0.81**	**0.70**	-0.03
GNM	**0.42**	**0.50**	0.10	**0.42**	**0.51**	0.11
HW2008 100	**0.64**	0.06	**0.45**	**0.60**	**0.37**	**0.40**
HW2008 200	**0.63**	0.06	**0.54**	**0.70**	**0.31**	**0.49**
H2011 UG	0.14	0.01	0.25			
SSP Only	**0.48**	**0.43**	**0.54**			

- N-grams good at memorizing known combinations in known context
  - Don't generalize well by context or by similarity to novel combinations
- Similarity and context necessary for generalizing to novel combinations
  - Hayes & Wilson 2008, Daland et al. 2011, Albright 2009, Jarosz & Rysling 2016

BALANCING FIT VS. GENERALIZATION

Challenges

- Identifying models that generalize like humans to unseen combinations
- Need to generalize to ‘similar’ patterns (a balance)
- Quantitative models (e.g. Bayesian, MDL, regularization) formalize this balance

Discoveries

- Capturing human generalization requires richer representations
- Results due to quantitative comparisons among quantitative models

Other competing pressures!

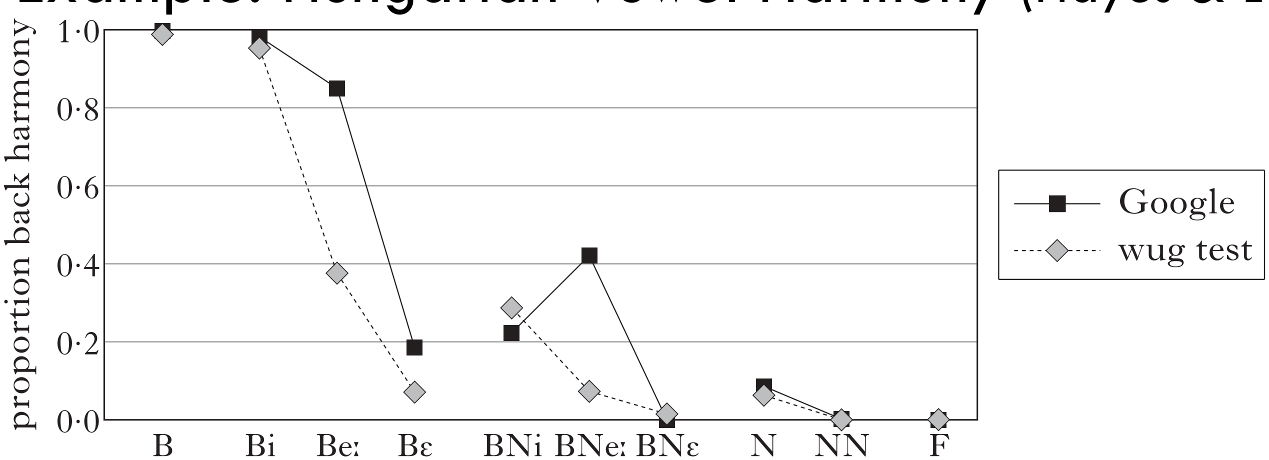
- Later: fit data or universal pressures

EMBRACING AMBIGUITY II: PATTERNED EXCEPTIONALITY

Patterned Exceptionality (Lexicalized Variation)

- Learners extend statistical trends to novel forms gradiently
- Individual words/morphemes exhibit fixed behavior
  - Zuraw 2000, 2010, Ernestus & Baayen 2003, Hayes & Londe 2006, Becker et al 2011, Gouskova & Becker 2013,
- Lexical propensities (Linzen et al. 2013, Jurgec 2016, Zymet 2018)

Example: Hungarian Vowel Harmony (Hayes & Londe 2006)



EMBRACING AMBIGUITY II: PATTERNED EXCEPTIONALITY

Modeling Challenges

- Learners treat known and novel items qualitatively differently
- Requires quantitative sensitivity
- Decision about where patterns should be attributed

Decisions & Trade-offs

- Should pattern be attributed to grammar or lexicon?
- Which data should each component explain?
- How do we ensure models generalize at all?

GENERALIZING FROM EXCEPTIONS: THREE HYPOTHESES

Threshold

- Regularization (e.g. Hudson Kam & Newport 2005), Past tense debate (e.g. Pinker & Prince 1988)
- Yang's Tolerance Principle (2016)

Frequency Matching

- Gradient Phonotactics, Lexicalized and Free Variation
- Proposed as a 'Law' in Hayes et al. (2009)
- Most work on lexicalized variation manually enforces this assumption (cf Zymet 2018)

Soft Threshold

- Generalizations in experiments are often skewed toward majority pattern
- Predictions of MaxEnt models of exceptionality learning (Moore-Cantwell & Pater 2016, Hughto et al 2019)

GENERALIZING FROM EXCEPTIONS: THREE HYPOTHESES

Ambiguity: should learner attribute pattern to grammar or lexicon?

- How are these components balanced?

Threshold

- Grammar for regular pattern
- Lexicon/Memorization for (limited number of) exceptions

Frequency Matching

- Grammar & Lexicon for all

Soft Threshold

- Mixture
- ‘Regular’ pattern more strongly encoded in grammar
- ‘Exceptions’ more memorized, less strongly encoded in grammar

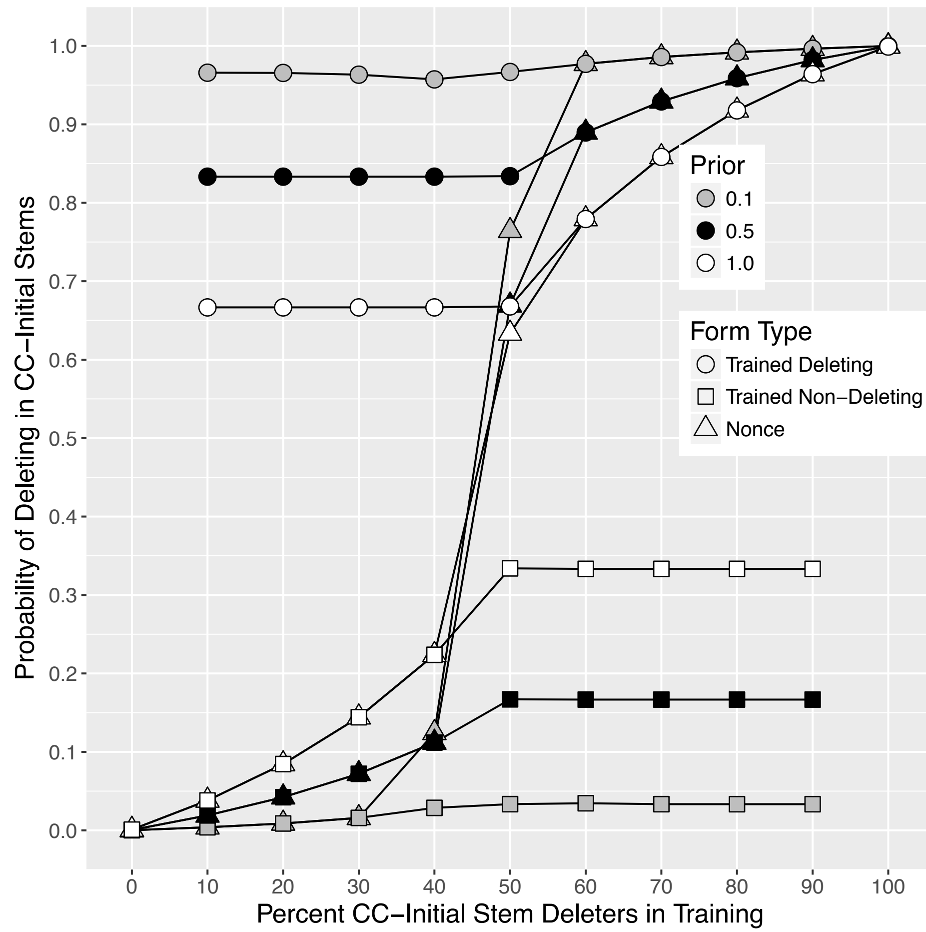
A CONTINUUM OF HYPOTHESES

Concrete Predictions

- Threshold (step function)
- Frequency Matching (y = x)
- Soft Threshold (pictured)

Takeaways

- All hypotheses are quantitative
- Quantitative modeling is required to compare these hypotheses
- Connection to quantitative behavioral data is required
- Testing human learners' generalization from incomplete data



Soft Threshold: Hughto, Lamont, Prickett, & Jarosz 2019

EMBRACING AMBIGUITY III: HIDDEN STRUCTURE

Quantitative modeling is useful for learning of categorical patterns with hidden structure.

Why?

- Quantitative models make it possible to formalize learners' gradient preferences among hypotheses
- Gradient preferences enable accumulation of information despite uncertainty

NOTHING IS CERTAIN

Learning datum: [tɛˈlɛfɔn]

Ambiguous

- Right-aligned Trochee: [tɛ(ˈlɛfɔn)]
- Left-aligned Iamb: [(tɛˈlɛ)fɔn]

But this is not lack of information!

- Prior beliefs: Right-aligned > Left-aligned
  - Trochaic > Iambic
- Prior beliefs: Left-aligned > Right-aligned
  - Iambic > Trochaic

Preferences among categorical hypotheses are gradient

- The stronger the prior beliefs, the stronger the inferences

Knowledge accumulates despite uncertainty

EMBRACING AMBIGUITY III: HIDDEN STRUCTURE

Extending quantitative machine learning methods to hidden linguistic structure has led to more successful, more robust learning models of

- Prosodic structure with constraints and parameters
  - Tesar & Smolensky 1998, 2000, Jarosz 2013, 2015, 2016, Boersma & Pater 2016, Nazarov & Jarosz 2017, Jarosz & Nazarov 2019
- Underlying representations
  - Jarosz 2006, 2015, Pater et al. 2012, Cotterell et al. 2015, Rasin & Katzir 2016
- Derivations
  - Jarosz 2016, Staubs & Pater 2016, Nazarov & Pater 2017, Rasin et al. 2018
- Exceptionality
  - Nazarov 2016, Moore-Cantwell & Pater 2016, Hughto et al 2019
- Rules & Constraints
  - Hayes & Wilson 2008, Calamaro & Jarosz 2015, Rasin et al. 2015, Rasin & Katzir 2016, Wilson & Gallagher 2018
- Hidden syntactic structure with constraints and parameters
  - Joint work in progress

RESULTS EXAMPLE: CONSTRAINTS

Applying principles of statistical inference to error-driven learning of constraint grammars

- More successful learners (Jarosz 2013)
- More efficient learners (Jarosz 2016)

Algorithm	Learning Rate (plasticity)			
	.05	.10	.25	.50
RIP/GLA	55.81 (1.82)	56.13 (1.62)	56.21 (2.15)	**57.50** (2.28)
RIP/SGA	**88.79** (0.97)	88.71 (0.66)	85.48 (1.57)	82.90 (2.92)
RRIP/GLA	**84.19** (1.91)	82.58 (1.91)	81.13 (2.29)	80.08 (3.09)
RRIP/SGA	**89.44** (0.71)	89.27 (0.94)	87.58 (1.79)	82.98 (1.84)
EIP/GLA	93.87 (0.78)	**93.95 (0.57)**	93.71 (1.69)	92.82 (1.29)
EIP/SGA	88.23 (0.56)	**88.31 (1.02)**	85.56 (1.96)	83.23 (2.57)

RESULTS EXAMPLE: PARAMETERS

Applying principles of statistical inference to learning of parameter setting (Nazarov & Jarosz 2017, Jarosz & Nazarov 2019)

> 90%

< 10%

	EDPL	NPL, no batch	NPL, batch = 5	NPL, batch = 10	Random baseline
# of runs that converge (% of 2800)	2644 (94.4%)	21 (0.8%)	176 (6.3%)	148 (5.3%)	
# of stress systems that converge at ≥1 run (% of 280)	268 (95.7%)	3 (1.1%)	25 (8.9%)	24 (8.6%)	
# of stress systems that converge at all 10 runs (% of 280)	255 (91.1%)	2 (0.7%)	10 (3.6%)	12 (4.3%)	
Median # of iterations/data points till convergence (range)	200 (100–15,700)	200,000 (4,400–9,999,900)	70,000 (400–9,000,000)	4,100 (700–9,999,900)	700 (100-30,000)

Faster than baseline

Slower than baseline

EMBRACING AMBIGUITY RECAP

Generalizations are gradient even for categorical data

With quantitative models we can

- formalize balance of competing pressures
- evaluate quantitative hypotheses on quantitative data
- accumulate knowledge despite uncertainty

Quantitative modeling is essential

- <u>Phenomena</u>: Coverage of quantitative phenomena
- <u>Computation</u>: Solutions to inconsistency and hidden structure learning challenges
- <u>Data</u>: Connecting to quantitative corpus and behavioral data
- <u>Evaluation</u>: Evaluating hypotheses on quantitative data

PARADIGM SHIFT: GRADIENCE IN LEARNING

- Shifting the role of learning in linguistic theory
  - Not (just) about what is or isn't (categorically) learnable
    - Gradient preferences among learnable patterns
  - Not (just) about what is or isn't representable
    - Gradient preferences among representable patterns
- Gradience handles choices <u>among representable patterns</u>
  - Novel connections to quantitative corpus and behavioral data
  - Novel methods and discoveries about gradient learning biases
  - Reframing connections between learning and UG

PARADIGM SHIFT: GRADIENCE IN LEARNING

Implications of Gradient Learn(ing|ability)

- Representable patterns can be harder/easier or faster/slower to learn
- Quantitative properties of the data affect learnability
  - Inherent learning biases to any quantitative learning model
- Soft learning biases interact with other soft biases

Learners don't perfectly reproduce their input

- They generalize some patterns more than others
- They skew: Under/over learn patterns relative to the input

Implications for language change, typology, and linguistic theory

- Detangle soft learning biases from grammatical pressures

DETANGLING SOFT LEARNING BIAS

Transparent and Opaque Derivations (Jarosz 2016)

- Some categorical patterns learned more quickly than others
- Quantitative modeling indicates <u>learning biases ***might*** derive observed skews</u>

Universal SSP in gradient phonotactics (Jarosz 2017, Jarosz & Rysling 2017, Jarosz & Rysling in prep)

- SSP is a soft bias that interacts with experience gradiently
- Quantitative modeling indicates <u>learning biases ***cannot*** derive observed skews</u>

LEARNING PROCESS INTERACTIONS
(JAROSZ 2016)

Which rule interactions are more 'natural'?

- Maximal utilization (Kiparsky 1968)
  - **Feeding & counterbleeding ≻ bleeding & counterfeeding**
- Transparency (Kiparsky 1971)
  - **Bleeding & feeding ≻ counterbleeding & counterfeeding**

What principles underlie 'naturalness'?

- Simpler, unmarked (Kiparsky 1968, 1971)
- Surface Truth / Exceptionality (Kenstowicz & Kisseberth 1977)
- Paradigm Uniformity / Leveling (Kiparsky 1971, Kenstowicz & Kisseberth 1977, Kenstowicz 1996, Benua 1997, McCarthy 2005)
- Recoverability / Contrast Preservation / Semantic Transparency (Kaye 1974, 1975, Kisseberth 1976, Gussmann 1976, Kenstowicz & Kisseberth 1977, Donegan and Stampe 1979, Łubowicz 2003)

LEARNING PROCESS INTERACTIONS (JAROSZ 2016)

- Are these principles grammar internal (e.g. in UG)?
  - Kiparsky (1971: 614)
    - "The hypothesis which I want to propose is that opacity of rules adds to the cost of the grammar"
  - Kiparsky (1971: 581)
    - "If … are hard to learn, the theory will have to reflect this formally by making them expensive"
- Questions
  - Could these principles be derived?
  - Why inconsistencies?
    - Sometimes counterbleeding > bleeding
    - Sometimes bleeding > counterbleeding
    - Sometimes rule re-ordering
    - Sometimes rule loss

LEARNING PROCESS INTERACTIONS
(JAROSZ 2016)

- Modeling Process Interactions
  - A statistical learning model for Harmonic Serialism
  - Serial Markedness Reduction (SMR; Jarosz 2014)
- Minimal UG & Learning Assumptions
  - No ranking is more ‘marked’ or more ’complex’ than any other
    - Some rankings produce opaque, some transparent interactions
    - Constraints start out ‘tied’ – no initial bias toward any ranking
  - No paradigm uniformity, no contrast preservation in UG
  - Model is sensitive to frequency
    - learns frequent, less ambiguous patterns more quickly

LEARNING PROCESS INTERACTIONS
(JAROSZ 2016)

- Simple learning system
  - Two processes
    - V → ∅ / __V
    - s → ʃ / __i
  - Four possible **interactions**

	a. Deletion	b. Palatalization	c. Bleeding	d. Feeding
UR	/su-a /	/si/	/si-a/	/su-i/
Deletion	sa	—	sa	si
Palatalization	—	ʃi	—	ʃi
SR	[sa]	[ʃi]	[sa]	[ʃi]

	a. Deletion	b. Palatalization	c. Counterbleeding	d. Counterfeeding
UR	/su-a/	/si/	/si-a/	/su-i/
Palatalization	—	ʃi	ʃia	—
Deletion	sa	—	ʃa	si
SR	[sa]	[ʃi]	[ʃa]	[si]

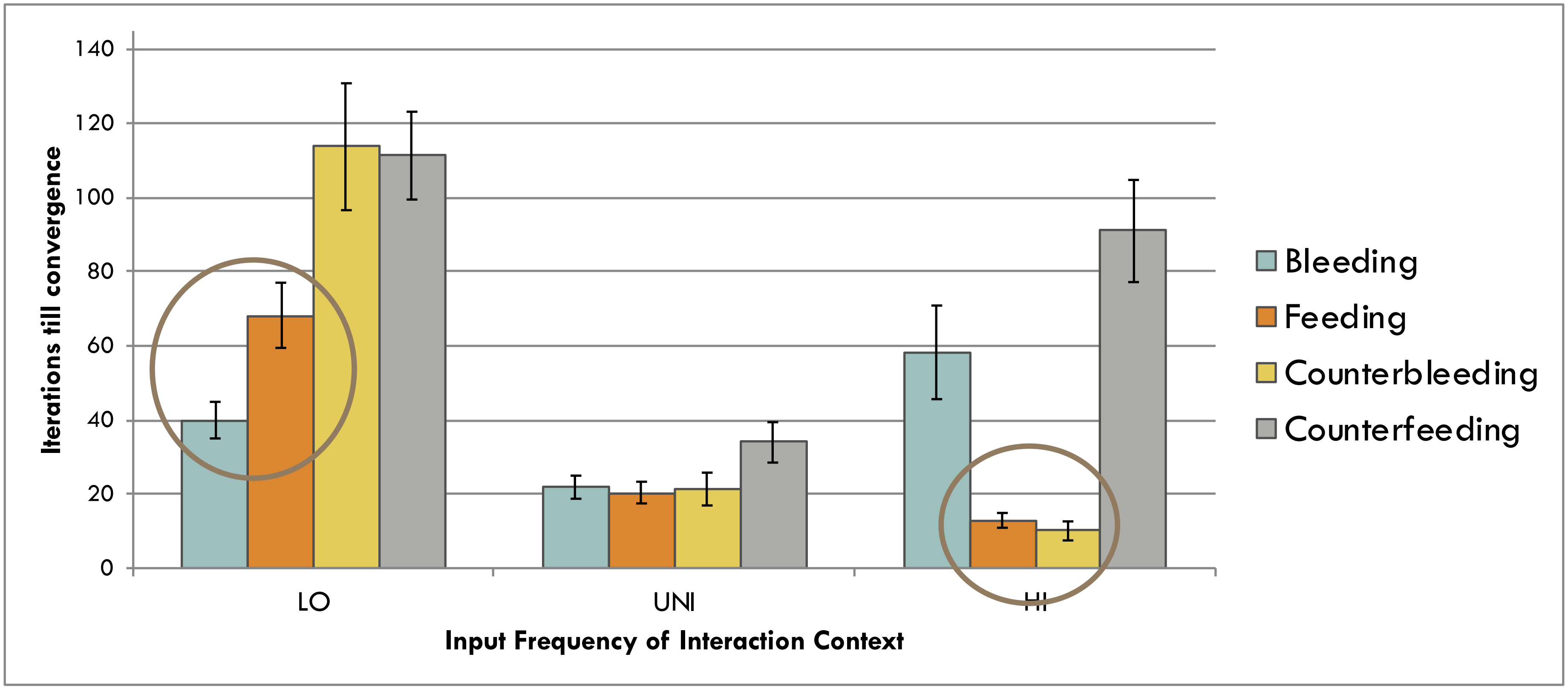
LEARNING PROCESS INTERACTIONS
(JAROSZ 2016)

- Four 'Languages' – 1 for each interaction
  - Deletion
  - Palatalization
  - **One interaction**
- Varied
  - Relative Frequency of interacting context **(HI, UNI, LO)**

	1 Bleeding	2 Feeding	3 Counterbleeding	4 Counterfeeding
Deletion	sua→sa	sua→sa	sua→sa	sua→sa
Palatalization	si→ʃi	si→ʃi	si→ʃi	si→ʃi
Interaction	**sia→sa**	**sai→ʃi**	**sia→ʃa**	**sai→si**
	lo uni hi	lo uni hi	lo uni hi	lo uni hi

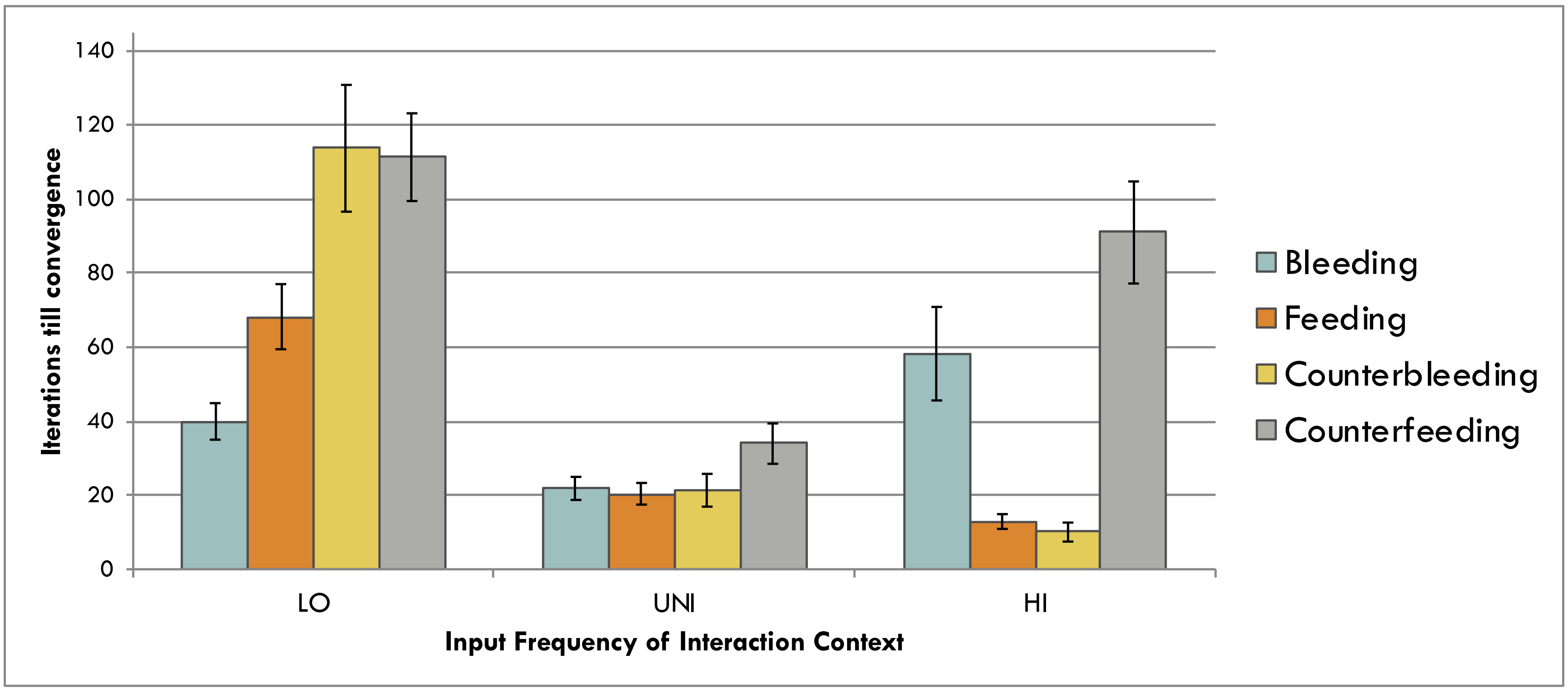
LEARNING INTERACTIONS
(JAROSZ 2016)

LO: transparent were easier to learn (Kiparsky 1971)
HI: maximally utilized were easier to learn (Kiparsky 1968)



LEARNING INTERACTIONS
(JAROSZ 2016)

LO: interaction is rare => learning of opaque interaction is slow
HI: palatalization is rare => learning of palatalization is slow



INTERACTIONS DISCUSSION

Basic UG + statistical learning => emergent biases

- More abundant & unambiguous evidence => faster learning

Detangle learning biases from UG

Predictions for human learning

- Prickett (2018): model predicts patterns in ALL experiment

Modeling connects UG and language change

- Novel prediction about effect of input frequency
- Novel prediction about re-ordering v. rule-loss
  - Transparency ⇔ re-ordering
  - Maximal utilization ⇔ rule loss

SSP IN GRADIENT PHONOTACTICS

Sonority Sequencing Principle (SSP; Clements 1988, Selkirk 1984)

[lb]ack ≺ [nb]ack ≺ [bd]ack ≺ [bn]ack ≺ [bl]ack ≺ [bj]ack

-2 -1 0 1 2 3

Consistent findings of **Sonority Projection** in English

- Preferences between unattested clusters
  - **#nb** (-1) vs. **#db** (0)

Documented using various tasks

- 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)

ENGLISH: NATURE OR NURTURE?

Berent et al. (2007): Nature

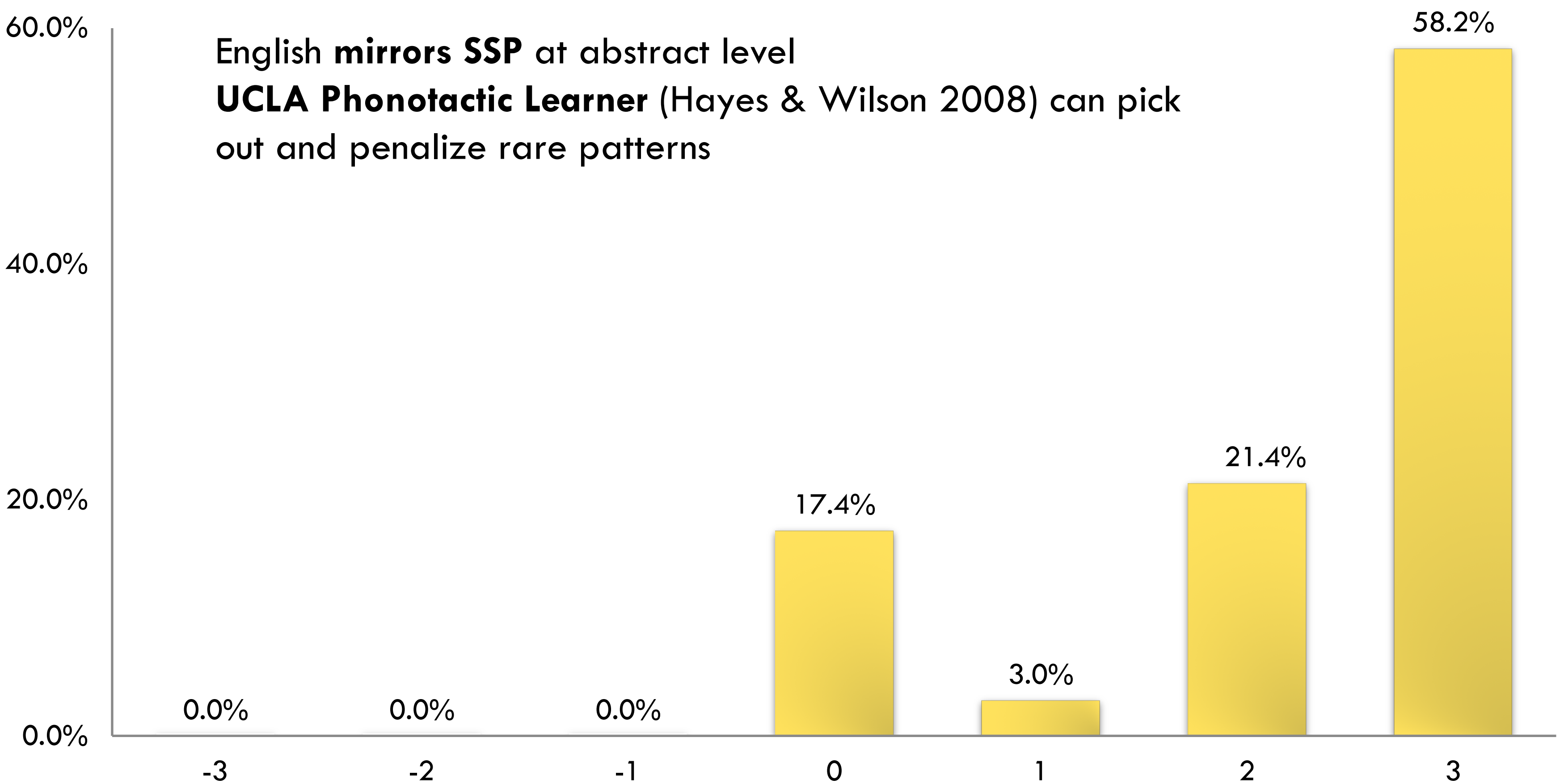
- English speakers exhibit sonority projection effects
  - *[lb]ack (-2) $\prec$ *[bd]ack (-1) $\prec$ *[bn]ack (1)
- Basic lexical statistics don't capture effect

Daland et al. (2011): Nurture

- models derive SSP for English (e.g. UCLA Phonotactic Learner Hayes & Wilson 2008)
- As long as statistical learning has access to
  - **Syllable structure** - [gb] in rug.by may be different
  - **Features** - what sounds are similar to one another
    - #bn similar to #sn, #bl, …
    - #nb much farther from #na, #sp
- With the right representations, **SSP may be derivable from statistics**

ENGLISH LEXICAL STATISTICS

English **mirrors SSP** at abstract level
UCLA Phonotactic Learner (Hayes & Wilson 2008) can pick out and penalize rare patterns



(data from Hayes & Wilson 2008)

English is not a strong test case

THE OPPOSITE?

Better test case: Polish?

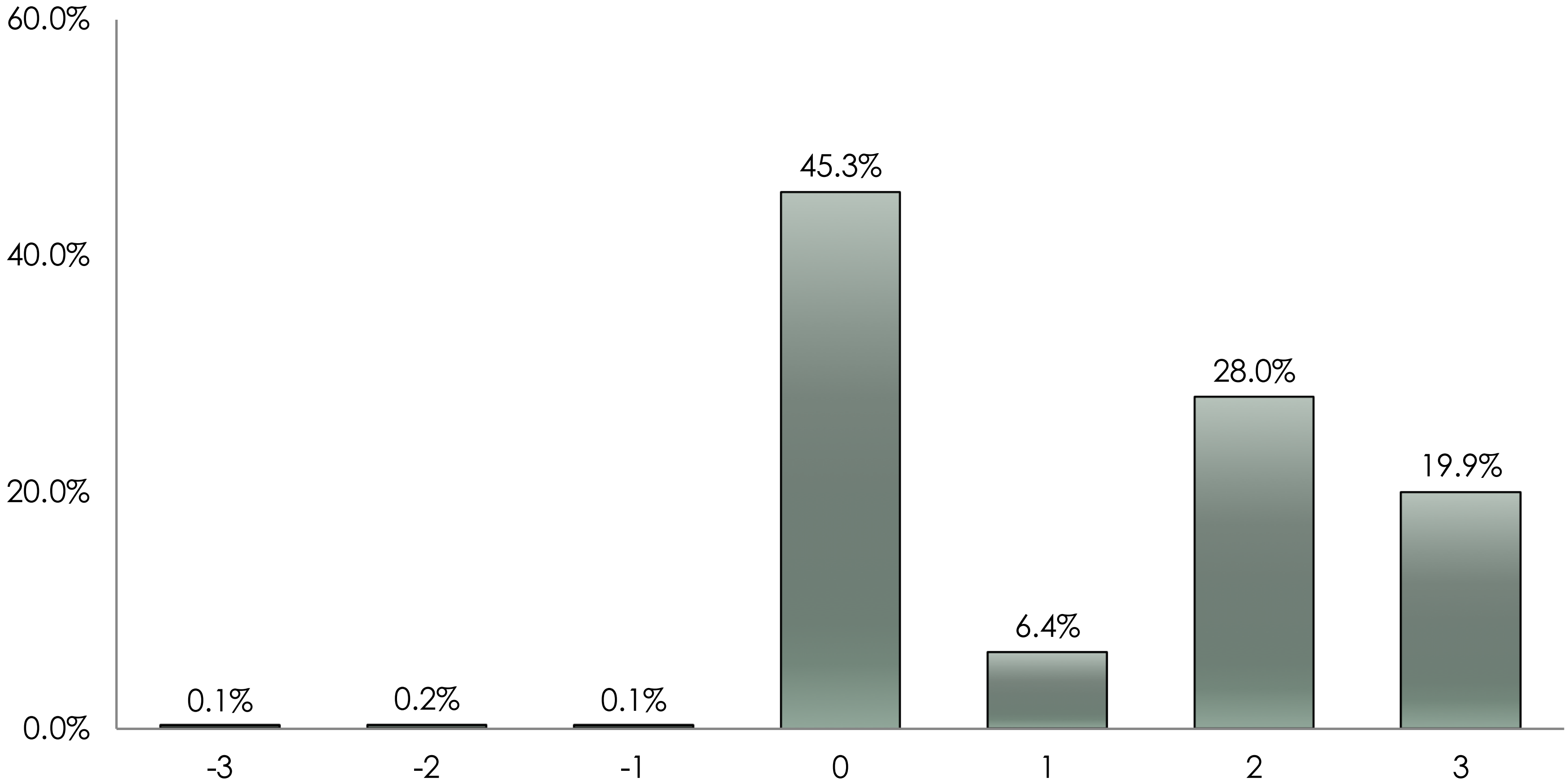
[wb]ack ≺ [lb]ack ≺ [mb]ack ≺ [bd]ack ≺ [bn]ack ≺ [bɹ]ack ≺ [bj]ack

-3	-2	-1	0	1	2	3
[wzɨ]	[lvɨ]	[mʂa]	[ptak]	[dnɔ]	[klutʃ]	[zwɨ]

What do the statistics look like?

- From Polish CDS Frequency Dictionary (Haman 2011)
  - ~800k word tokens (~115k #CC)
  - ~44k word types (~11k #CC)
- Numbers very look similar in text, inflectional dictionaries

POLISH LEXICAL STATISTICS



SSP SENSITIVITY IN POLISH?

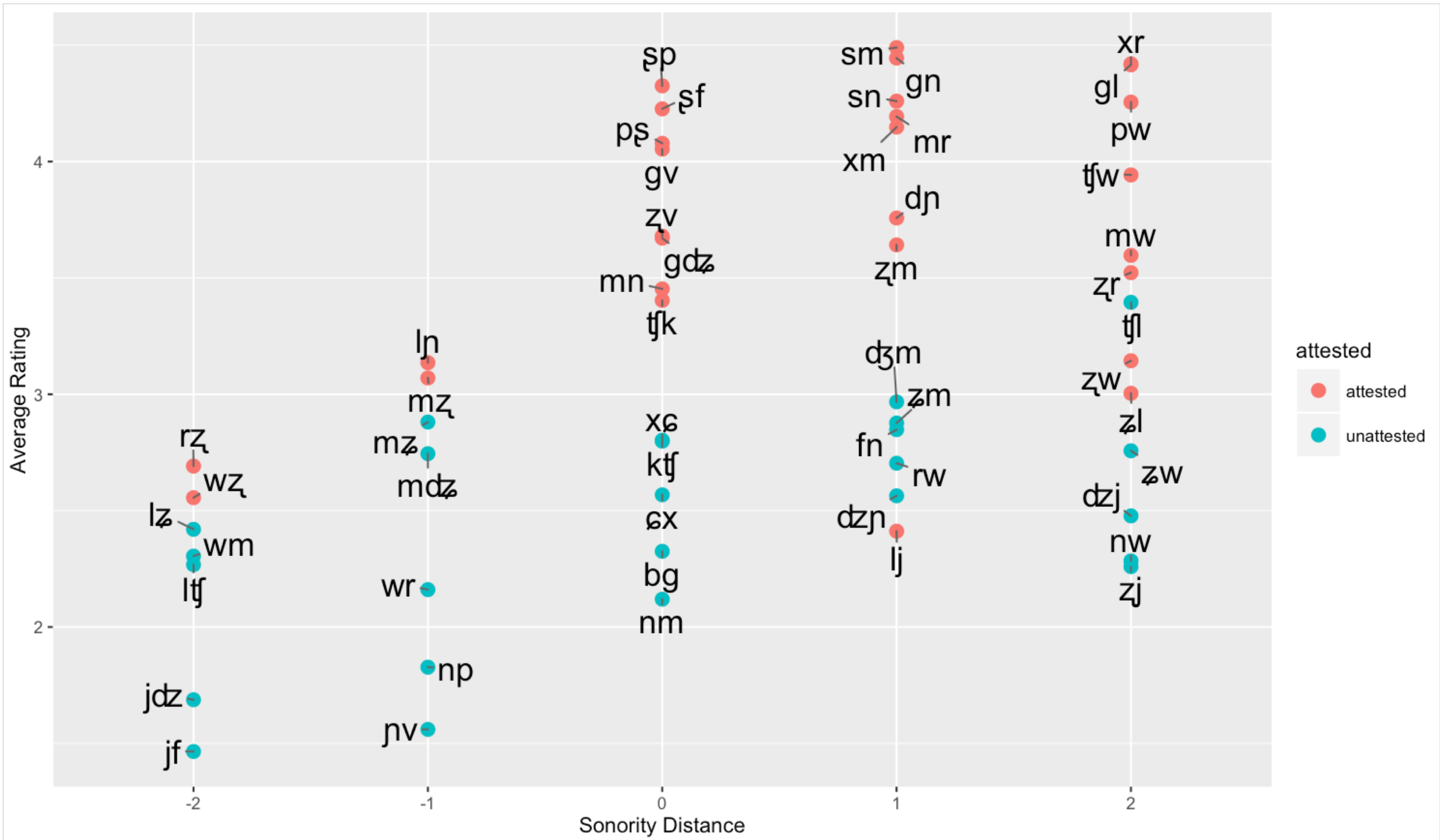
Previous Work

- Traditional analyses: SSP active in phonology
  - <u>Comparative Allomorphy</u> (Rubach 1986; Bethin 1987; Rubach & Booij 1990a, 1990b)
  - <u>Voicing Processes</u> (Rubach & Booij 1987, 1990, 1990b)
- Acquisition of Polish
  - Later development of sonority falls (Łukaszewicz 2006, 2007)
  - 1;7-2;6 yo more accurate on higher rises (Jarosz 2017)

Experiment (Jarosz & Rysling 2016)

- Are adults' phonotactic judgments driven by SSP?
  - Sonority Rise -2/3 thru +2/3
- How does generalization work?
  - Attested vs. Unattested clusters

RESULTS: AVERAGE RATINGS BY CLUSTER & ATTESTEDNESS



RESULTS

Ordinal mixed effects model

- Dependent: Rating
- Fixed effects: SSP * Attestedness
- Full Random FX, by Subject, Tail

Results

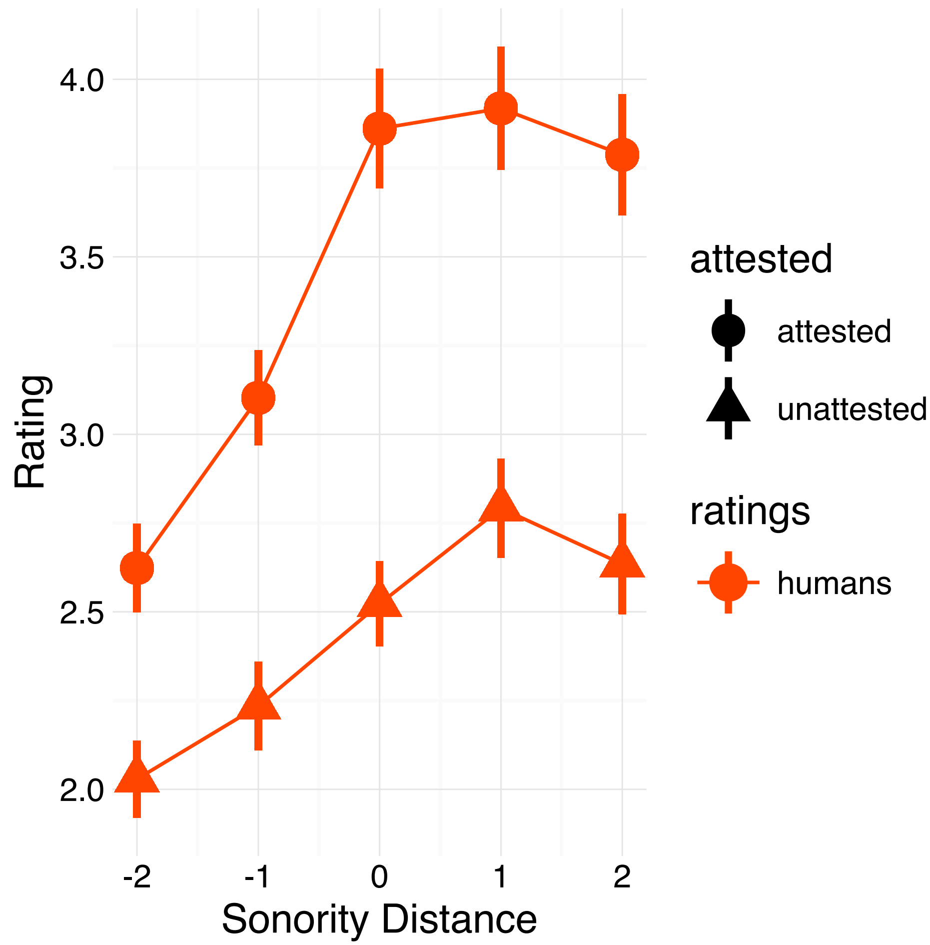
- **SSP** (**β=0.28**, z=9.38)
- **Attestedness** (**β=0.90**, z=17.48)
- interaction n.s. (β=-0.005, z=-0.30)

Overall SSP trend

- Same for attested and unattesteds

Jarosz & Rysling (in prep)

- Flattening/reversals are interactions with experience



MODELING OVERVIEW

Trained on phonetically transcribed Polish lexicon

- Derived from child directed speech to 1;6-3;2
- ~44k word types

Models from previous work

- Phoneme Bigram & Trigram
- Grapheme Bigram & Trigram
- Neighborhood/Analogical (GNM: Bailey & Hahn 2001)
- UCLA Phonotactic Learner (Hayes & Wilson 2008)
- UCLA Learner with Sonority UG (Hayes 2011)

Training (following Daland et al. 2011)

- Word transcriptions
- Syllabified word transcriptions
  - Maximal onset with observed word-initial clusters

MODELS FAIL TO CAPTURE SSP

	Unsyllabified	Syllabified
	SSP β (t)	*SSP β (t)*
Grapheme Bigram	**0.24 (10.52)**	
Grapheme Trigram	**0.20 (8.78)**	
Phoneme Bigram	**0.25 (10.65)**	**0.13 (5.67)**
Phoneme Trigram	**0.16 (7.34)**	**0.15 (7.22)**
GNM	**0.30 (13.31)**	**0.30 (13.31)**
HW2008 100	**0.23 (10.09)**	**0.19 (8.19)**
HW2008 200	**0.22 (9.71)**	**0.15 (6.53)**
H2011 UG	**0.23 (10.31)**	

- Do these models capture SSP effect in ratings?
  - Fit: ratings ~ model
  - Fit: residuals ~ SSP + (1+SSP | tail) + (1+SSP | subject)
    - Is there still effect of SSP after factoring out models' predictions?
  - Significant positive coefficient on SSP indicates failure to account for effect of SSP in ratings

SOFT SSP BIAS

Statistical learning with rich representations is insufficient

- No unbiased model captures **overall SSP trend** in both attesteds and unattesteds

Quantitative Modeling

- Unbiased/Unconstrained models fail: not derivable from learning
- Human learning is biased by SSP
- Bias is soft – interacts with experience
- Neither pressure is absolute

LEARNING BIASES DISCUSSION

Biases are soft, *quantitative skews*

Statistical learning automatically predicts skews/biases

Existing Progress & Discoveries

- Better learning performance
- Detangling learning biases and grammatical theory

But much more to be done!

NEXT DIRECTIONS

Quantitative Modeling + Hidden Structure + Corpus/Exp Data

- Models can do this now!
- Compare predictions of representationally rich theories on corpus data representative of linguistic experience and evaluate on experimental data learning and generalization
  - Provide novel sources of evidence for long-standing theoretical debates

Understanding implications of ambiguity, quantitative patterns for development, language change, and typology

- Information is gradient
- We need more exploration of how this affects learning rates and outcomes

CONCLUSIONS

Nothing is certain (and it's ok!)

- We (as scientists) know how to deal with it
- We (as language learners) know how to deal with it

Quantitative modeling

- Connects theory to quantitative corpus and behavioral data
- Connections -> discoveries about soft biases
- Progress on detangling of learning and other biases
- Still a lot we don't understand about inherent learning biases

THANK YOU

SELECTED REFERENCES

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26(01). 9–41.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44(4). 568–591.

Becker, M., A. Nevins & N. Ketrez. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in turkish laryngeal alternations. *Language* 87(1). 84–125.

Berent, Iris, Donca Steriade, Tracy Lennertz & Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. *Cognition* 104(3). 591–630.

Boersma, Paul & Joe Pater. 2016. Convergence Properties of a Gradual Learning Algorithm for Harmonic Grammar. In John McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Calamaro, Shira & Gaja Jarosz. 2015. Learning general phonological rules from distributional information: A computational model. *Cognitive science* 39(3). 647–666.

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. *arXiv preprint cmp-lg/9707017*.

Cotterell, Ryan, Nanyun Peng & Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association of Computational Linguistics* 3(1).

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann. 2011. Explaining sonority projection effects. *Phonology* 28(02). 197–234.

Ernestus, M. & R. H. Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 5–38.

SELECTED REFERENCES

Gouskova, Maria & Michael Becker. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory* 31(3). 735–765.

Hayes, B., K. Zuraw, P. Siptár & Z. Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4). 822–863.

Hayes, Bruce & Zsuzsa Cziraky Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23(01). 59–104. doi:10.1017/S0952675706000765.

Hayes, Bruce & Colin Wilson. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry* 39(3). 379–440. doi:10.1162/ling.2008.39.3.379.

Hudson Kam, Carla L. & Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59(1). 30–66. doi:10.1016/j.cogpsych.2009.01.001.

Hughto, Coral. 2018. Investigating the Consequences of Iterated Learning in Phonological Typology. *Proceedings of the Society for Computation in Linguistics*, vol. 1, 182–185. doi:10.7275/R5WH2N63. https://scholarworks.umass.edu/scil/vol1/iss1/21.

Hughto, Coral, Andrew Lamont, Brandon Prickett & Gaja Jarosz. 2019. Learning exceptionality and variation with lexically scaled MaxEnt. *Proceedings of the Society for Computation in Linguistics*, vol. 2, 91–101.

Jarosz, Gaja. 2006a. *Rich Lexicons and Restrictive Grammars - Maximum Likelihood Learning in Optimality Theory*. PhD Dissertation, the Johns Hopkins University, Baltimore, MD.

Jarosz, Gaja. 2006b. Richness of the Base and Probabilistic Unsupervised Learning in Optimality Theory. *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL*, 50–59. New York City, USA: Association for Computational Linguistics.

SELECTED REFERENCES

Jarosz, Gaja. 2009. Restrictiveness and Phonological Grammar and Lexicon Learning. In Malcolm Elliot, James Kirby, Osamu Sawada, Eleni Staraki & Suwon Yoon (eds.), *Proceedings of the 43rd Annual Meeting of the Chicago Linguistics Society*, vol. 43, 125–134. Chicago Linguistics Society.

Jarosz, Gaja. 2013. Learning with Hidden Structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology* 30(1). 27–71.

Jarosz, Gaja. 2014. Serial Markedness Reduction. In John Kingston, Claire Moore-Cantwell, Joe Pater & Robert D. Staubs (eds.), *Proceedings of the Annual Meetings on Phonology*, vol. 1. Linguistic Society of America.

Jarosz, Gaja. 2015. Expectation Driven Learning of Phonology. Manuscript. University of Massachusetts, Amherst, ms.

Jarosz, Gaja. 2016. Learning opaque and transparent interactions in Harmonic Serialism. *Proceedings of the Annual Meetings on Phonology*, vol. 3.

Jarosz, Gaja. 2017. Defying the stimulus: acquisition of complex onsets in Polish. *Phonology* 34(2). 269–298.

Jarosz, Gaja & Amanda Rysling. 2017. Sonority Sequencing in Polish: the Combined Roles of Prior Bias & Experience. *Proceedings of the Annual Meetings on Phonology* 4(0). doi:10.3765/amp.v4i0.3975 (8 October, 2017).

Kiparsky, Paul. 1968. Linguistic universals and linguistic change. In Bach, Emmon & Robert T. Harms (eds.), *Universals in linguistic theory*, 170–202. New York: Holt, Reinhart & Winston.

Kiparsky, Paul. 1971. Historical linguistics. In W. O. Dingwall (ed.), *A Survey of Linguistic Science*, 576–642. College Park: University of Maryland Linguistics Program.

Linzen, Tal, Sofya Kasyanenko & Maria Gouskova. 2013. Lexical and phonological variation in Russian prepositions. *Phonology* 30(3). 453–515.

SELECTED REFERENCES

Moore-Cantwell, Claire & Joe Pater. 2016. Gradient Exceptionality in Maximum Entropy Grammar with Lexically Specific Constraints. *Catalan Journal of Linguistics* 15(0). 53–66. doi:https://doi.org/10.5565/rev/catjl.183.

Moore-Cantwell, Claire & Robert D. Staubs. 2014. Modeling Morphological Subgeneralizations. *Proceedings of the Annual Meetings on Phonology* 1(1). doi:10.3765/amp.v1i1.42. https://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/42 (16 January, 2018).

Nazarov, Aleksei. 2016. Extending Hidden Structure Learning: Features, Opacity, and Exceptions. *Doctoral Dissertations*. http://scholarworks.umass.edu/dissertations_2/782.

Nazarov, Aleksei & Gaja Jarosz. 2017. Learning Parametric Stress without Domain-Specific Mechanisms. *Proceedings of the Annual Meetings on Phonology*, vol. 4. Washington, DC: Linguistic Society of America. (8 October, 2017).

Nazarov, Aleksei & Joe Pater. 2017. Learning opacity in Stratal Maximum Entropy Grammar. *Phonology* 34(2). 299–324.

Pater, Joe. 2012. Emergent systemic simplicity (and complexity). In J Loughran & A McKillen (eds.), *Proceedings from Phonology in the 21st Century: In Honour of Glyne Piggott. McGill Working Papers in Linguistics*, vol. 22.

Prickett, Brandon. 2018. Experimental evidence for biases in phonological rule interaction. Lisbon, Portugal.

Rasin, Ezer, Iddo Berger & Roni Katzir. 2015. Learning rule-based morpho-phonology. MIT, Cambridge, MA, ms.

Rasin, Ezer & Roni Katzir. 2016. On Evaluation Metrics in Optimality Theory. *Linguistic Inquiry* (47). 235–82.

Staubs, Robert D. & Joe Pater. 2016. Learning serial constraint-based grammars. In John J. McCarthy & Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

SELECTED REFERENCES

Tesar, Bruce & Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29(2). 229–268.

Vitevitch, Michael S. & Paul A. Luce. 2004. A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers* 36(3). 481–487. doi:10.3758/BF03195594.

Wilson, Colin & Gillian Gallagher. 2018. Accidental gaps and surface-based phonotactic learning: a case study of South Bolivian Quechua. (49). 610–23.

Yang, Charles. 2016. The price of productivity. *Manuscript, University of Pennsylvania.*

Zuraw, Kie. 2000. *Patterned Exceptions in Phonology*. UCLA.

Zymet, Jesse. 2018. *Lexical propensities in phonology: corpus and experimental evidence, grammar, and learning*. UCLA PhD Thesis.