REFINING UG: CONNECTING PHONOLOGICAL THEORY AND LEARNING

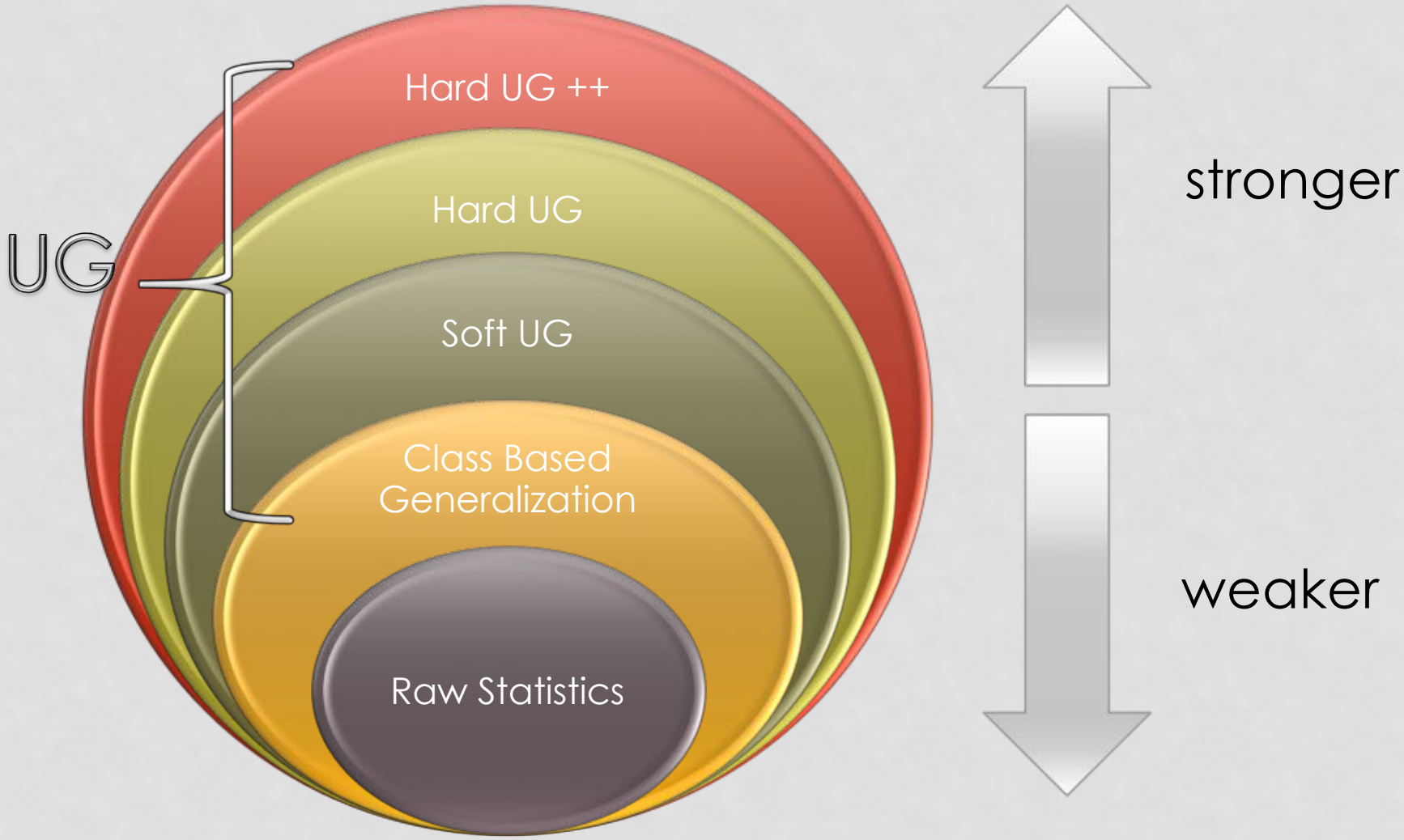
GAJA JAROSZ

NELS 47
OCTOBER 14-16, 2016, UMASS, AMHERST, MA

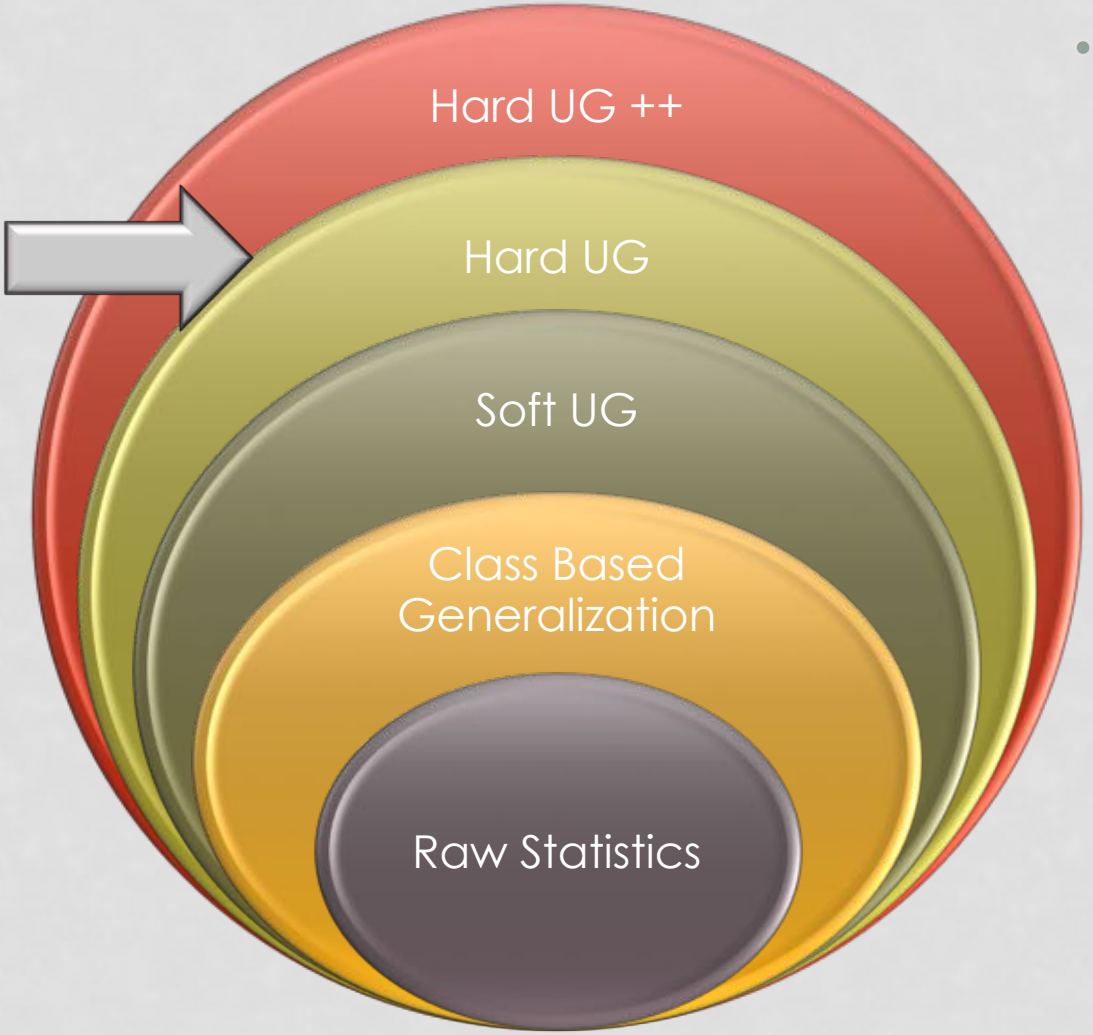
REFINING UG/LAD

- Universal Grammar vs. Statistics
  - False dichotomy
  - Universal Grammar + Statistics (c.f. Yang 2004)
    - Human learning is statistical
  - Continuum of hypotheses
- Computational modeling helps refine/define
  - How strong is UG?
  - What's the nature of UG?
- This talk
  - Overview of the continuum and themes
  - Three case studies illustrating this approach

AN LAD CONTINUUM

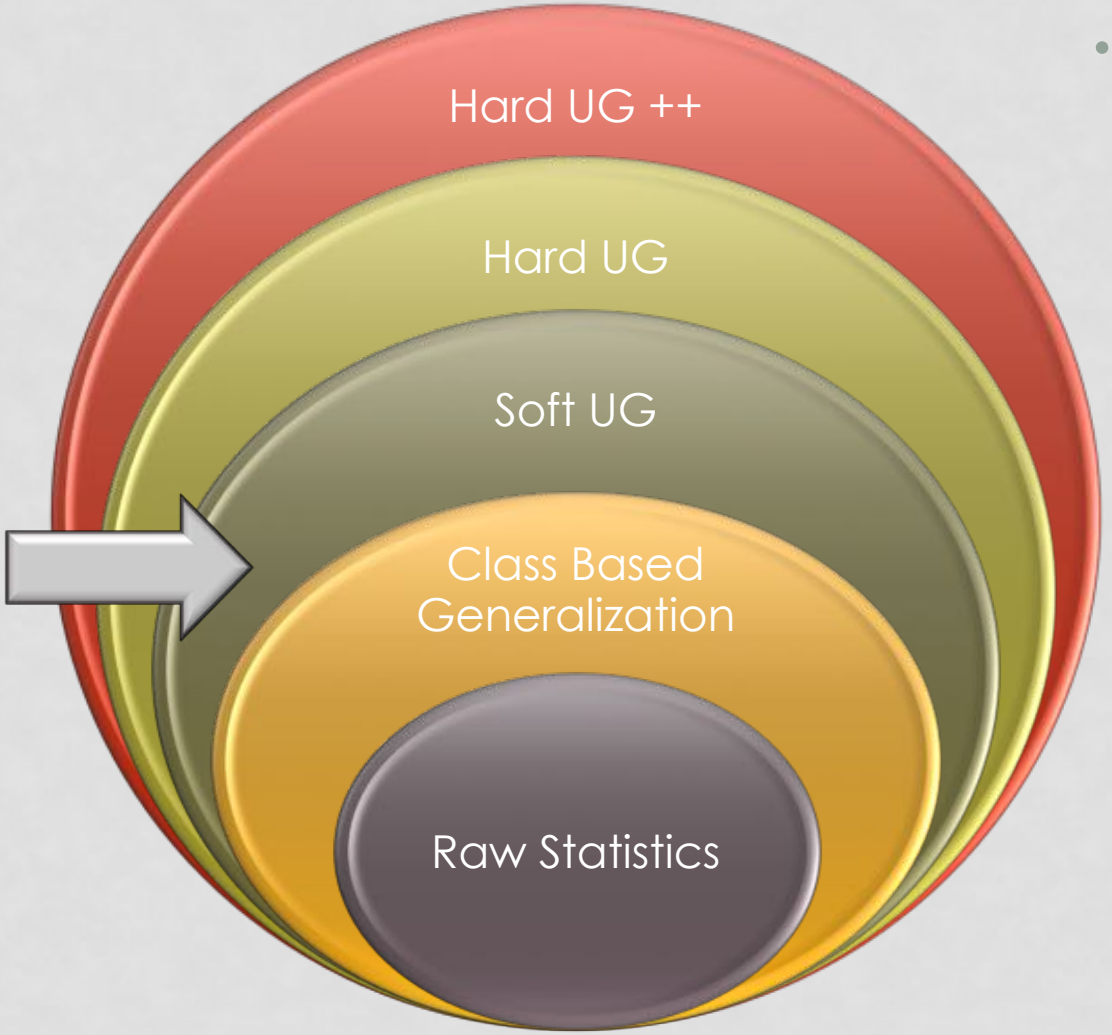


AN LAD CONTINUUM



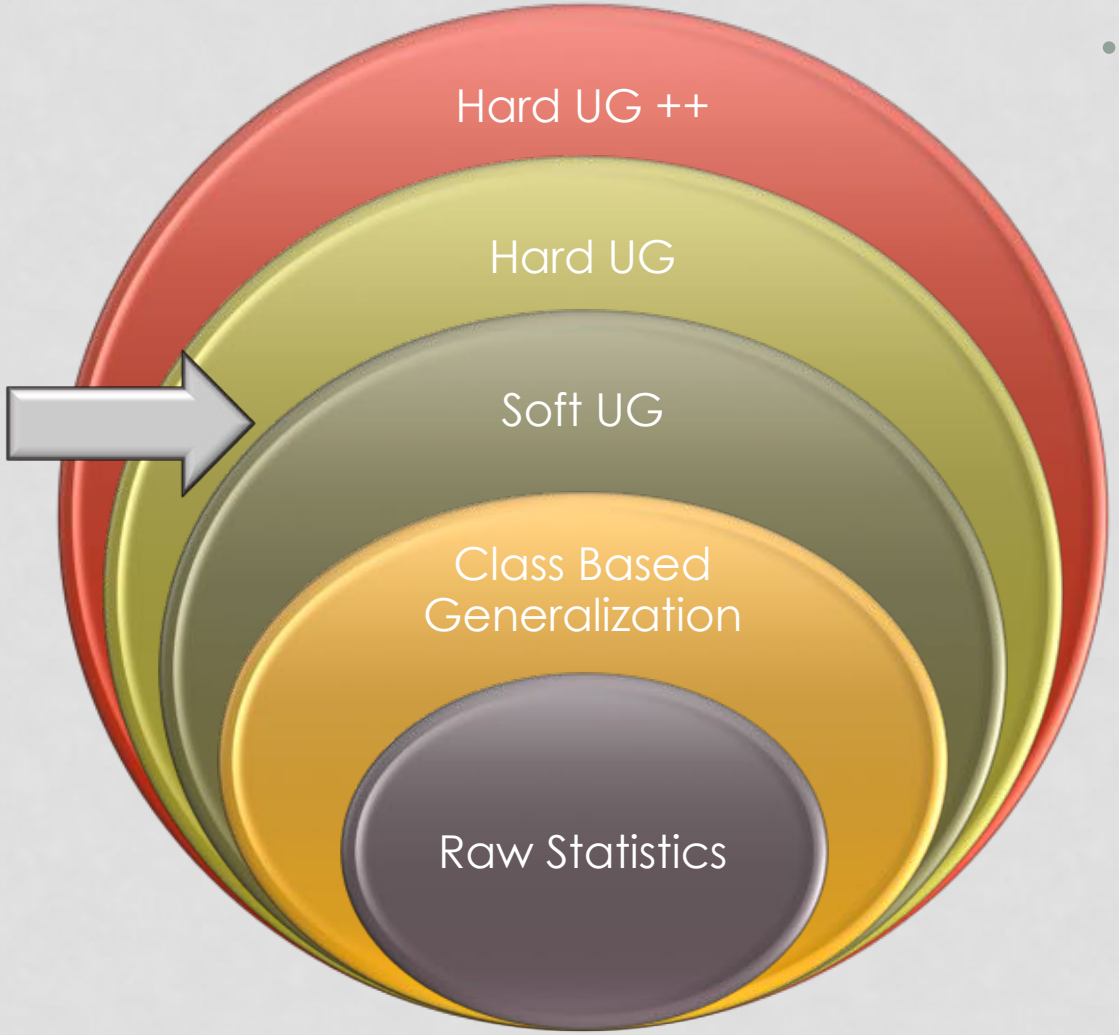
- Traditional UG
  - Hard ≠ non-probabilistic
  - Some patterns are categorically impossible
  - Hard cut between
    - Possible languages
    - Impossible languages
  - Example
    - Classic OT (Prince & Smolensky 1993) or Stochastic OT (Boersma 1997) with **universal constraints**
    - *VoicedCoda
    - mo**p** > mo**b**

AN LAD CONTINUUM



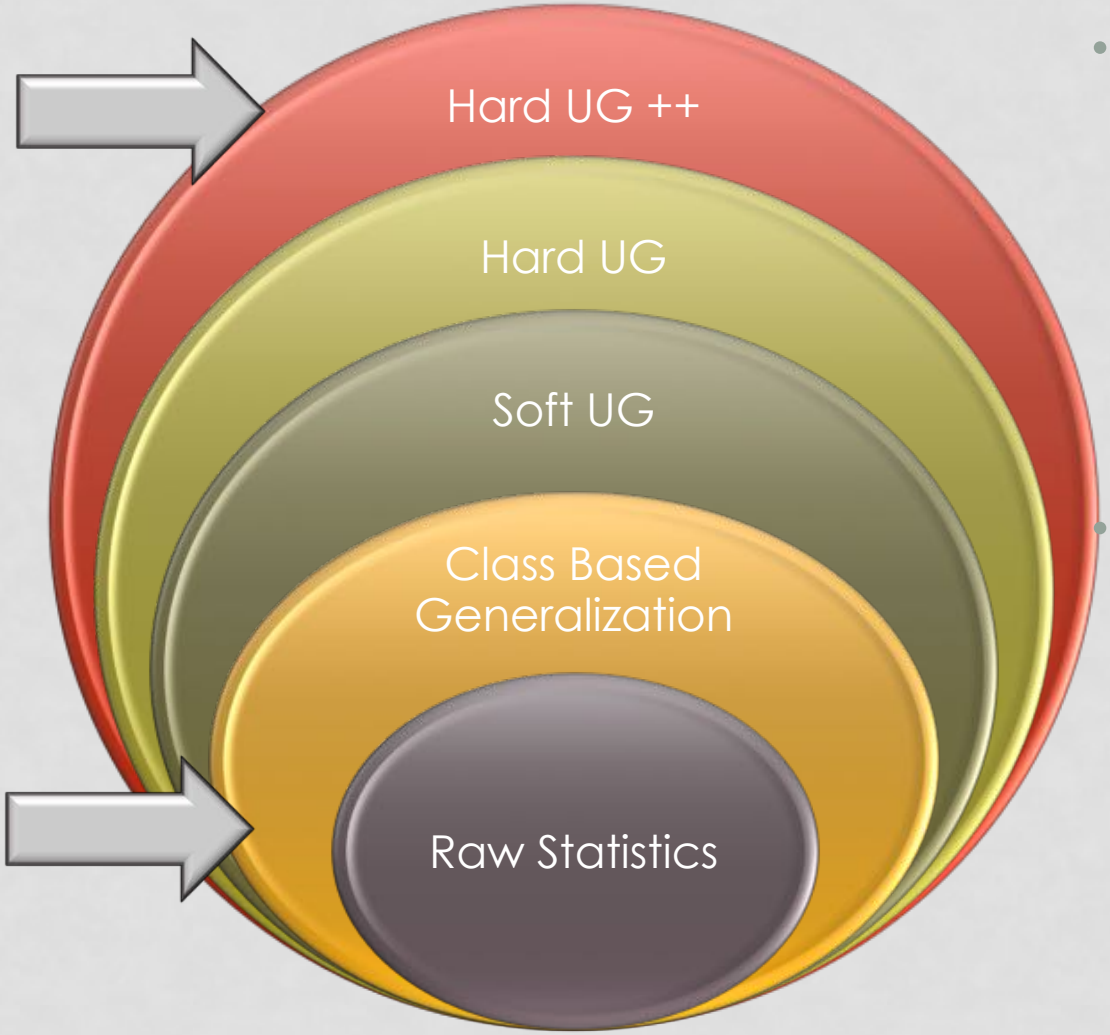
- Just representations
  - UG provides
    - Features
    - Syllables
    - Feet
    - Tiers
    - …
  - But learning over these representations is driven entirely by the input
    - UCLA Phonotactic Learner (Hayes & Wilson 2008)
    - *Voice**less**Coda
    - mo**b** > mo**p**

AN LAD CONTINUUM



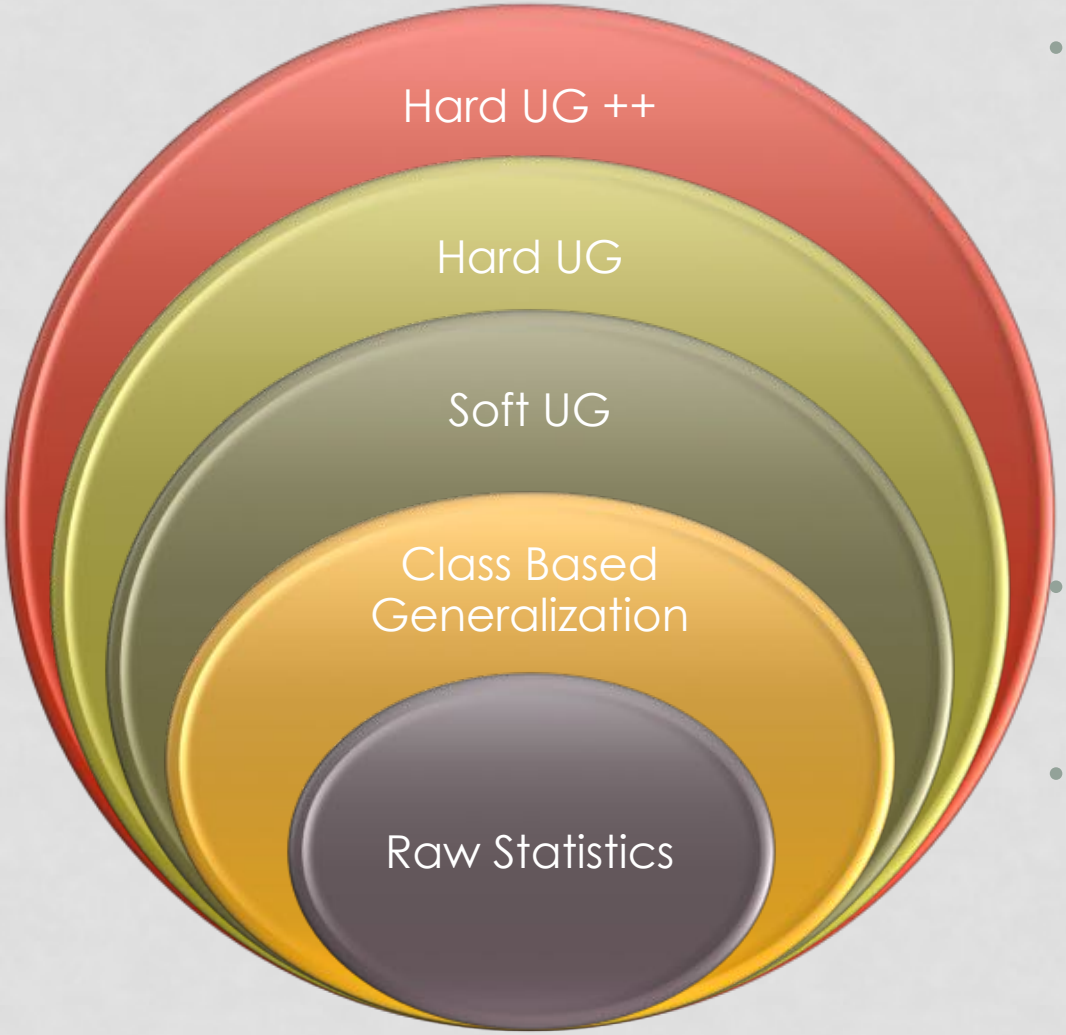
- Representations + soft bias
  - Balances built-in biases and input statistics
  - Harder to learn some patterns than others
    - Wilson (2006), Culbertson et al. (2012, 2013)
    - *Voice**less**Coda
    - mo**b** > mo**p**

AN LAD CONTINUUM



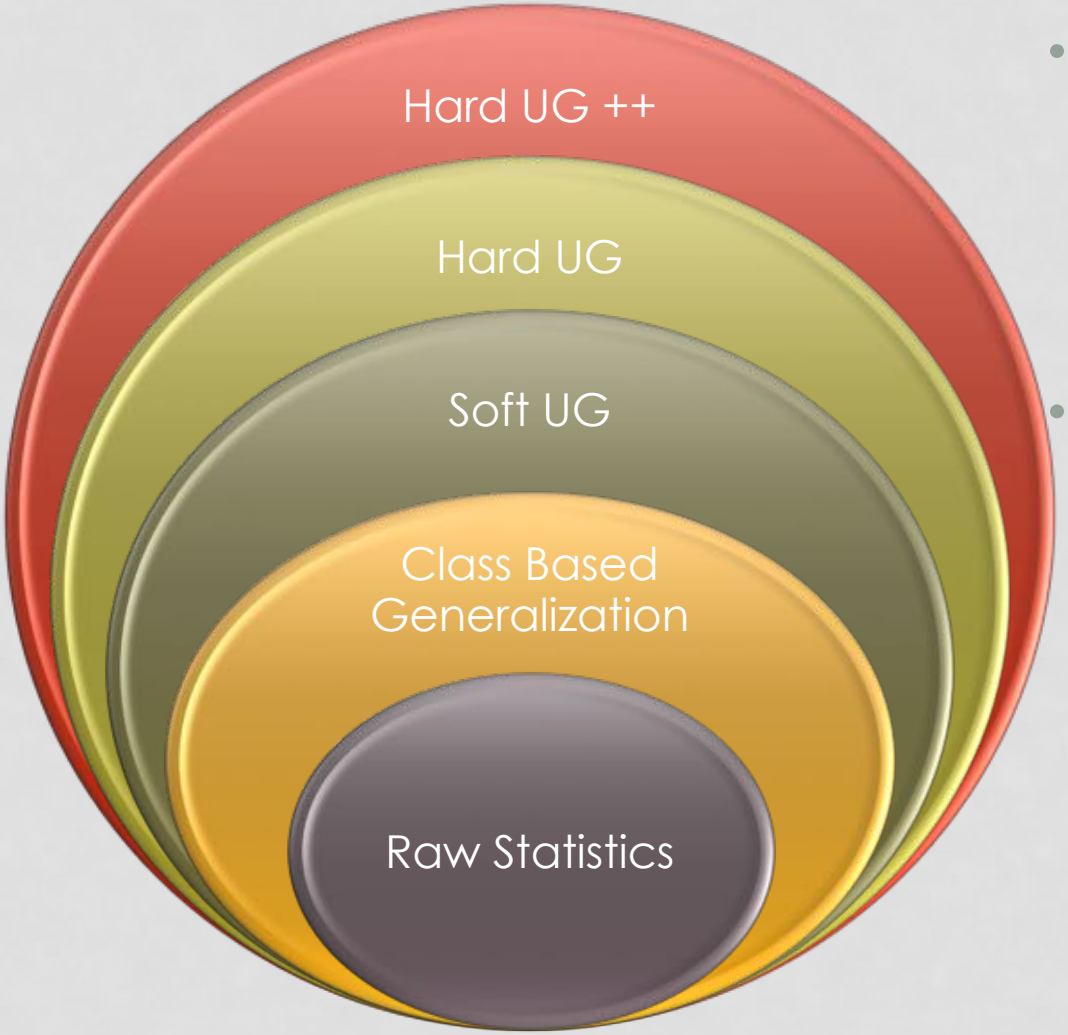
- Even Stronger
  - Hard UG + domain-specific learning
    - Substantive cues for parameter setting (Dresher & Kaye 1990)
    - Learning biases specific to language M >> F (Gnanadesikan 1995)
- Weakest
  - No UG, just statistics
  - Example
    - Basic analogical models
    - N-grams

RECENT THEMES



- Disentangling Universals
  - Universal ⇒ UG ?
    - Maybe not
  - UG embedded in more complex system
  - Maybe some universals can be derived from other properties of the system
- Modeling Learning
  - Bottom-up Inductionism
  - Top-down Reductionism
- Refinement
  - Differences across domains
  - Identify classes of universals

TALK OVERVIEW



- Three Examples
  - Rule/Process Order
  - Learning Parameter Settings
  - Syllable Phonotactics
- Evidence for both
  - weakening UG
  - strengthening UG

LEARNING INTERACTIONS (JAROSZ 2016)

- Which rule interactions are more 'natural'?
  - Maximal utilization (Kiparsky 1968)
    - **Feeding & counterbleeding > bleeding & counterfeeding**
    - Also Anderson (1969, 1974)
  - Transparency (Kiparsky 1971)
    - **Bleeding & feeding > counterbleeding & counterfeeding**
    - Also Kaye (1974, 1975), Kenstowicz & Kisseberth (1977)
- 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 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"
- Question
  - Could these principles be derived?
    - Could the learning difficulty follow from the kinds of patterns these interactions present to the learner?
  - Why inconsistencies and indeterminancies?
    - Sometimes counterbleeding > bleeding
    - Sometimes bleeding > counterbleeding
    - Sometimes rule re-ordering
    - Sometimes rule loss

LEARNING INTERACTIONS
(JAROSZ 2016)

- Modeling Process Interactions
  - A statistical learning model for Harmonic Serialism
  - Serial Markedness Reduction (SMR; Jarosz 2015)
    - SM constraints can favor opaque derivations
    - Ranked just like other constraints
- 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 patterns more quickly

LEARNING 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]

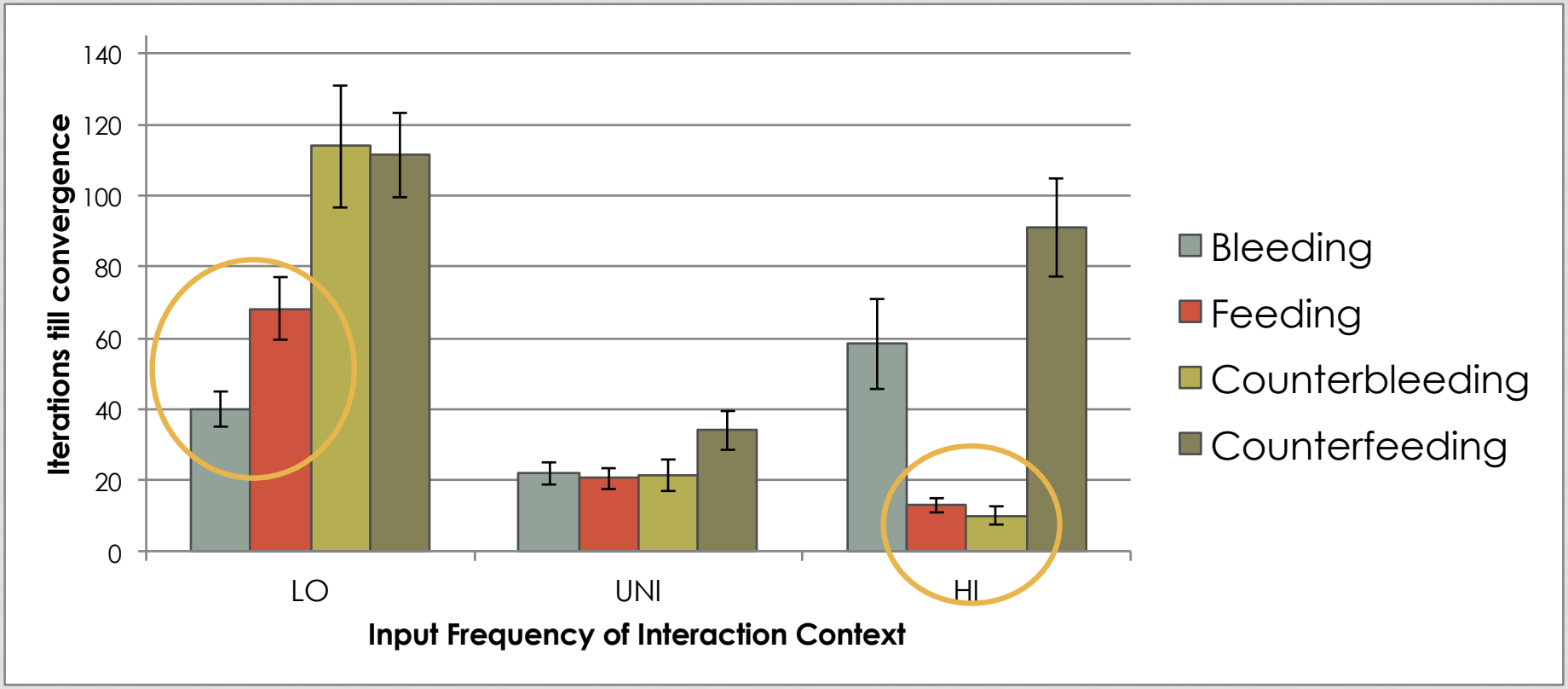
LEARNING INTERACTIONS
(JAROSZ 2016)

- 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

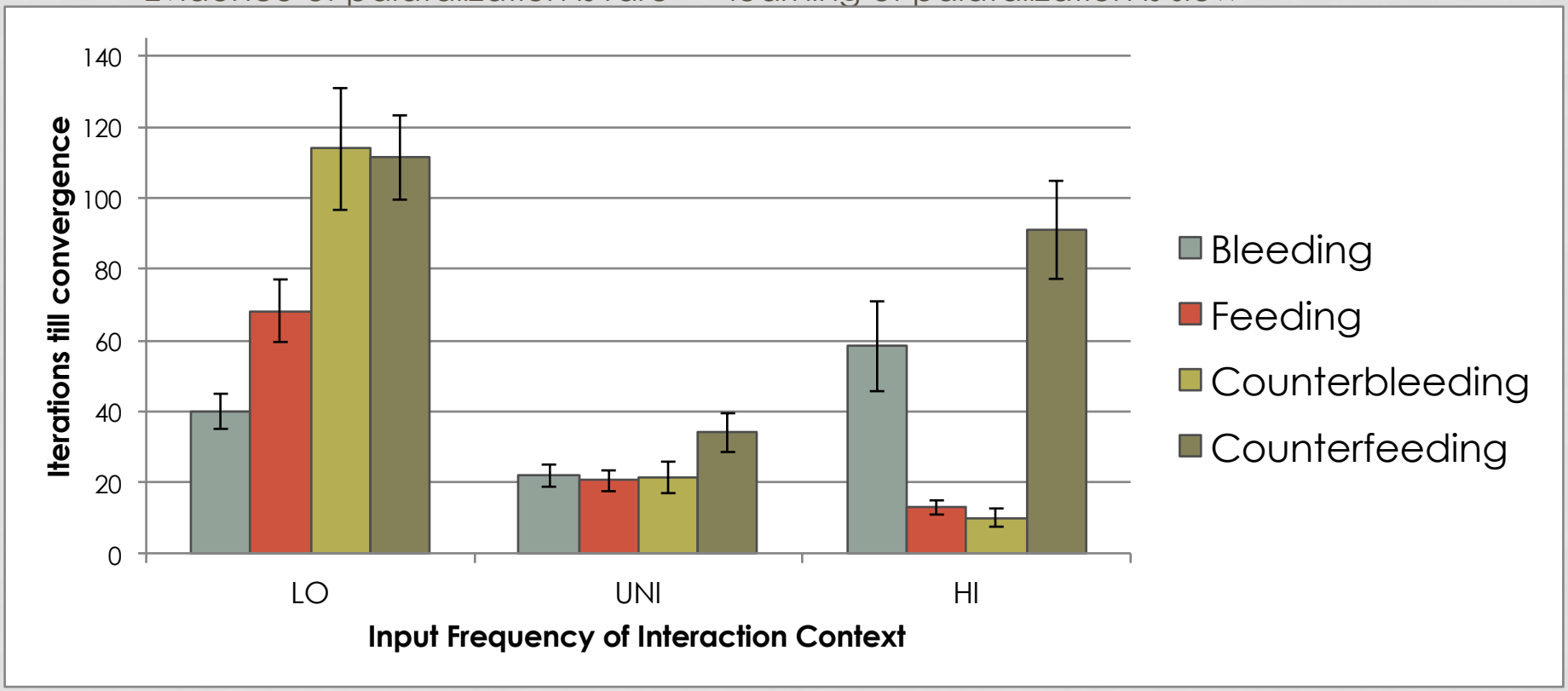
LEARNING INTERACTIONS
(JAROSZ 2016)

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



LEARNING INTERACTIONS
(JAROSZ 2016)

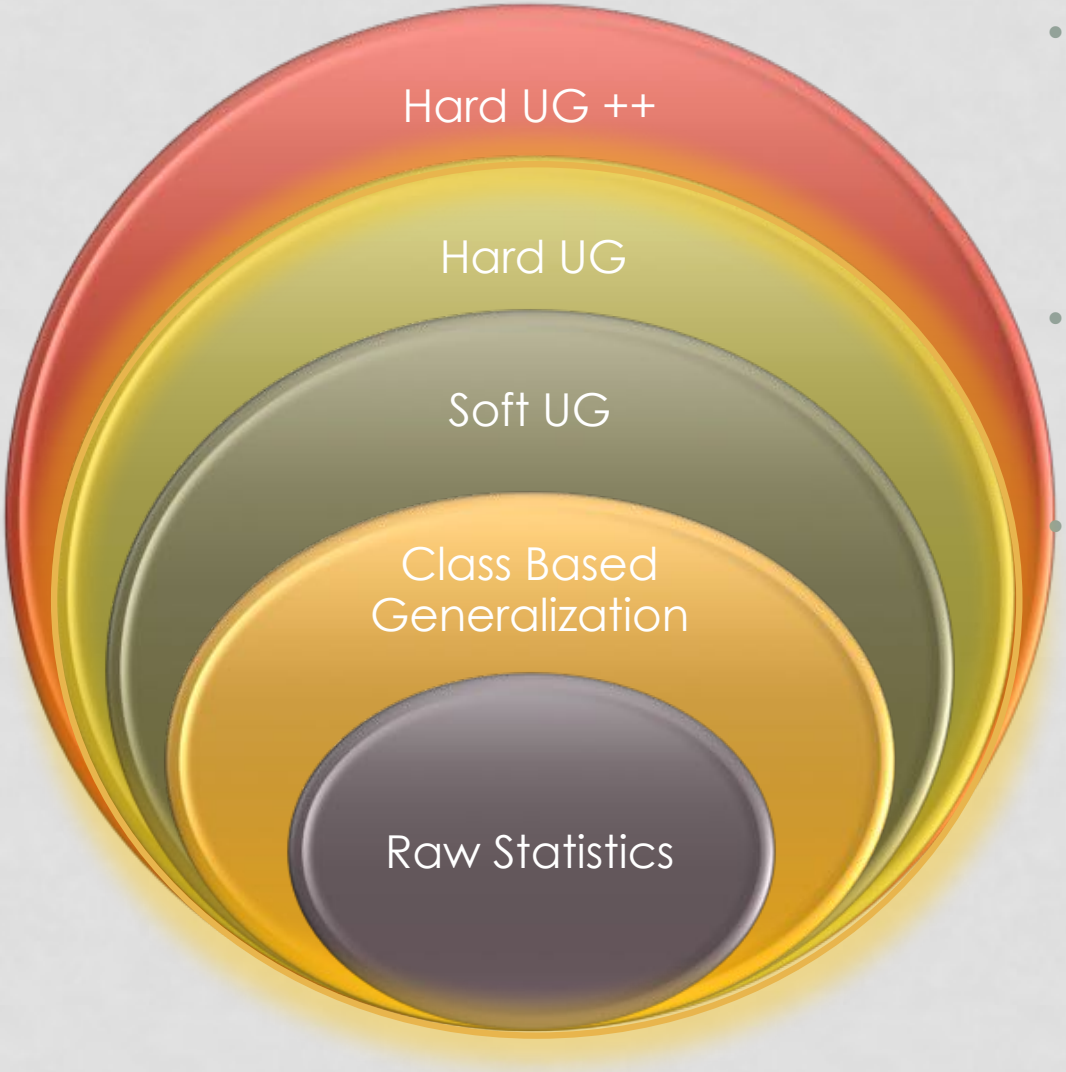
- LO: very particular ranking needed for opaque interaction
  - Evidence of interaction is rare => learning of opaque interaction is slow
- HI: F and CB provide evidence of both processes, B/CF just deletion
  - Evidence of palatalization is rare => learning of palatalization is slow



LEARNING INTERACTIONS
(JAROSZ 2016)

- Basic UG + statistical learning => emergent biases
- The nature of the learning problem varies
  - Inherent learning difficulties to some patterns
    - Basic learning principle: more abundant & explicit evidence => faster learning
  - Opaque => requires more 'fine-tuning' of ranking
  - Maximal utilization => provides more abundant evidence of both processes
- These biases are automatic, unavoidable, and interacting
  - Even in this most basic starting model
- Novel, testable predictions
  - Novel prediction about how CB could be preferred
    - Not paradigm uniformity, no contrast preservation
    - Abundance of evidence in the input
  - Novel prediction about effect of input frequency
  - Novel prediction about re-ordering v. rule-loss
    - Transparency ⇔ re-ordering
    - Maximal utilization ⇔ rule loss

LEARNING INTERACTIONS: DISCUSSION



- Summary
  - Opaque rankings don't 'cost' more
    - Harder to learn anyway
  - No contrast preservation, no paradigm uniformity
    - CB can have an advantage
- Testable predictions
  - If wrong or insufficient
    - Refine and extend
  - Provides concrete hypotheses for Hard UG
- Deriving Universals
  - Complexity
    - Pater (2012), Moreton & Pater (2012), Pater & Moreton (2012), Moreton et al (2015)
  - Distributional Skews
    - Staubs (2015), Stanton (2016)
  - Compositionality
    - Smith et al. (2016), Culbertson & Kirby (2016)

LEARNING PARAMETERS
(NAZAROV & JAROSZ 2016 / IN PREP)

- Chomsky (1981): Principles and Parameters
  - UG: universals (principles) and finite choices (parameters)
  - Language learner's task: find settings of parameters
- Applied to stress systems (Dresher and Kaye 1990; Hayes 1995)
  - UG
    - Directionality: L-to-R or R-to-L?
    - Foot form: Trochee or Iamb?
    - Extrametricality: Yes: Right
  - Learning σ ˌσ σ ˈσ σ
    - (σ ˌσ)(σ ˈσ) σ ⇒ Left, Iambs…
    - σ (ˌσ σ)(ˈσ σ) ⇒ Right, Trochees…
    - <σ> (ˌσ σ)(ˈσ σ) ⇒ Extrametricality left, Trochees…

PREVIOUS PROPOSALS: HARD UG ++

- Previous proposals for stress: Hard UG ++
- Dresher and Kaye (1990) (see also Dresher 1994)
  - Parameters set in innately specified order
  - Each parameter has default value
  - Each parameter innately associated with a "cue"
    - Configuration in data that triggers marked value
    - E.g., QS starts out set to Off. If corpus contains two words of same length with different stress, set QS to On.
- See similar work on "triggering" in learning syntax (Gibson and Wexler 1994, Berwick and Niyogi 1996, Lightfoot 1999)

PREVIOUS PROPOSALS: HARD UG ++

- Previous proposals for stress: Hard UG ++
- Pearl (2007, 2011)
  - Statistical P&P, still Hard UG
  - Naïve Parameter Learner (NPL; Yang 2002)
    - Domain-general learner for parameters (syntax or phonology)
  - NPL does not succeed on English stress
  - Hard UG fails ⇒ Hard UG ++
  - ++: Domain-specific learning mechanisms are necessary
    - Parameter ordering
    - Cues (Dresher and Kaye) or parsing method (Fodor 1998, Sakas and Fodor 2001) for disambiguation

OUR PROPOSAL

- Enriched statistical learning model (Hard UG)
  - Expectation Driven Parameter Learner (EDPL; based on Jarosz submitted)
    - Works exactly like NPL, except
    - More nuanced update, individual to each parameter
- Domain general EDPL and NPL tested on languages predicted by Dresher and Kaye (1990)
  - First typologically extensive tests for NPL
  - EDPL massively outperforms NPL
- We argue that conclusions about necessity of domain-specific mechanisms are premature
  - Failure of NPL is not representative of all domain-general models

NPL & EDPL OVERVIEW

NPL	EDPL
Stochastic parameter grammar	
Grammar incrementally updated by Linear Reward-Penalty Scheme (Bush and Mosteller 1951) after each data point	

• Test all parameter settings ***simultaneously*** (one time)
  • Match → reward all parameters equally
  • Mismatch → penalize all parameters equally
	• Test each parameter ***individually*** (fixed # of times)
	  • Reward more successful settings more
	  • Computation still linear in the number of parameters

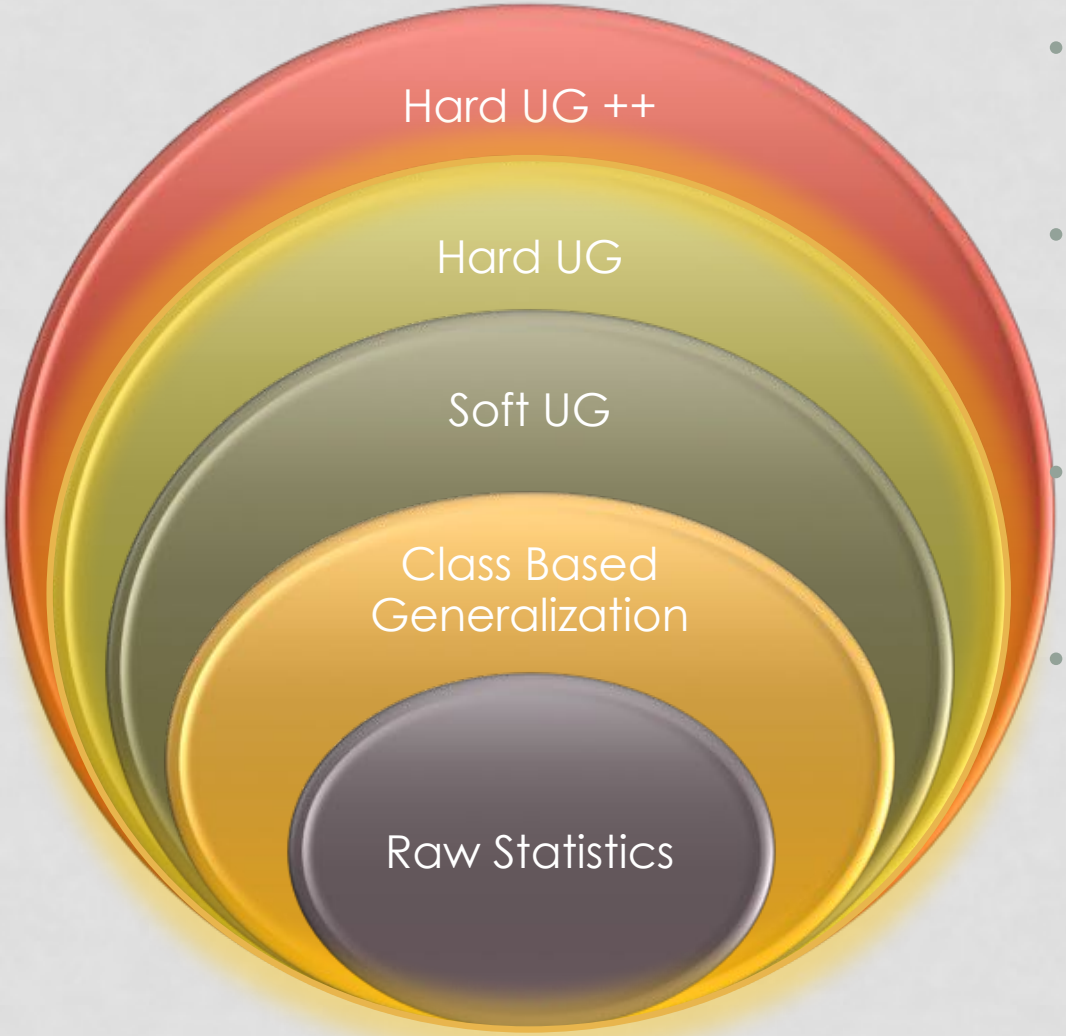
NPL V. EDPL

- Testing determines 'reward' amount $\boldsymbol{R(\psi_i)}$ for each parameter
  - $p(\psi_i)_{new} = \boldsymbol{R(\psi_i)} * \lambda + p(\psi_i)_{old} * (1 - \lambda)$
- NPL: $\boldsymbol{R(\psi_i)}$ is 0 or 1, <u>for all</u> selected parameters
  - 0 if combination of parameters doesn't match
  - 1 if combination of parameters does match
  - No ability to differentiate relevance of parameters
- EDPL: $\boldsymbol{R(\psi_i)}$ is a probability between 0 and 1, individually tested for each parameter
  - $\boldsymbol{R(\psi_i)}$ is 1 if $\boldsymbol{\psi_i}$ is crucial for that data point
  - $\boldsymbol{R(\psi_i)}$ is 0 if $\boldsymbol{\psi_i}$ is incompatible with that data point
  - $\boldsymbol{R(\psi_i)}$ is .5 if $\boldsymbol{\psi_i}$ is irrelevant to that data point
  - Each update increases the probability of successful analyses
    - $p(\psi_i|data\ point, G)$

STRESS TYPOLOGY TESTS

- Simulations
  - 23 Languages in Dresher and Kaye's system (1990)
    - Focus on 6 out of 10 parameters
  - No biases: all parameters start out unset
  - Each model tested 10 times on each language
- Results
  - NPL failed to converge for all but one language
    - Overall success rate: **4.3%**
    - within **89,370** iterations on average
  - EDPL showed convergence for all languages
    - Overall success rate: **96.0%**
    - within **200** iterations on average

EXAMPLE 2: DISCUSSION

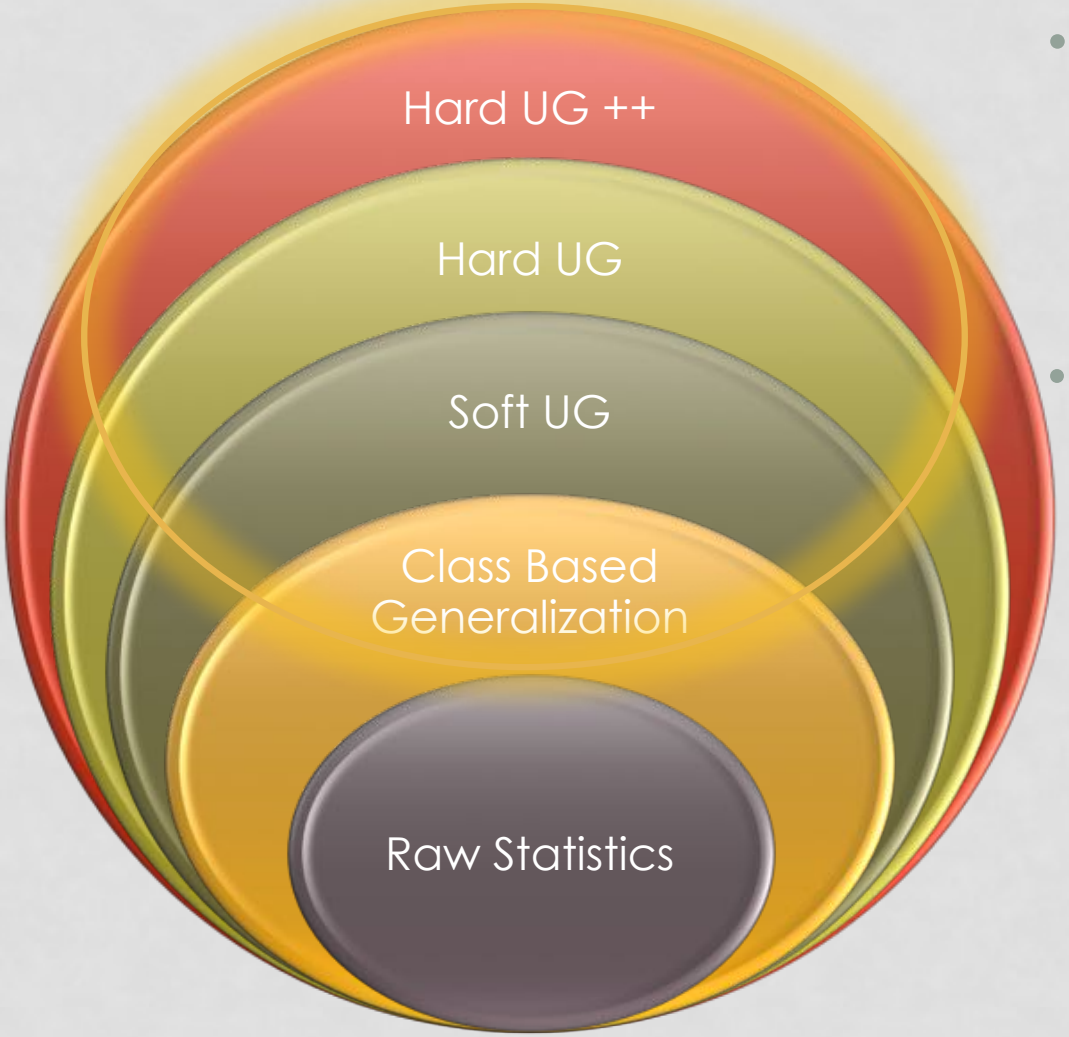


- Proposal: Hard UG +
  - basic statistical inference
  - Initial promising results
- Conclusions about Hard UG++ premature
  - Based on ineffective statistical learning model
- Next: push this model
  - See how far this can go
  - Investigate other domains
- Too soon to give up
  - Such strong and fragile assumptions about LAD may not be necessary

EXAMPLE 3: LEARNING PHONOTACTICS

- Where do gradient phonotactic preferences come from?
  - 'mip' > 'bwip' > 'dlap' > 'bzap'
- Lexicalist Hypothesis: Phonotactics derive from lexical statistics
  - Dominant hypothesis in phonology & language acquisition
  - Modeling work often assumes this
- UG: Universal principles constrain possible syllables & phonotactics
  - Modeling work has not emphasized UG

BOTTOM-UP INDUCTIONISM



- Modeling
  - Bottom-up inductionism
  - Evidence for strengthening LAD assumptions
- Frequency Sensitivity with Class-Based Generalization (CBG)
  - abstract representations: features, syllables, tiers, etc.
  - UCLA Phonotactic Learner (Hayes & Wilson 2008)
  - Featural Bigram Model (Albright 2009)

SUPPORT FOR CBG

- UCLA Phonotactic Learner (Hayes & Wilson 2008)
  - English onsets phonotactic judgments (from Scholes 1966)

Model	*r*
Our model	0.946
Clements and Keyser 1983 constraints with maxent weights	0.936
Coleman and Pierrehumbert 1997	0.893
Our model without features	0.885
N-gram model	0.877
Analogical model	0.833

- This is a lexicalist model:
  - It constructs constraints for under-represented patterns
    - *[+son,+dor] - no dorsal nasals
    - *[+son][ ] - no sonorant-initial clusters
  - Weights constraints to match lexical patterns
- This is a CBG model: it uses features & natural classes:
  - See also Albright (2009), Daland et al. (2011), Coetzee & Pater (2008), Albright & Hayes (2003)

SONORITY PROJECTION

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

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

-2 -1 0 1 2 3

- Consistent findings of **Sonority Projection** in English
  - Preferences between unobserved 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)
- Question:
  - Where do these preferences come from?
    - Statistics are 0 for all
  - Could they be learned?

ENGLISH: POVERTY OF THE STIMULUS?

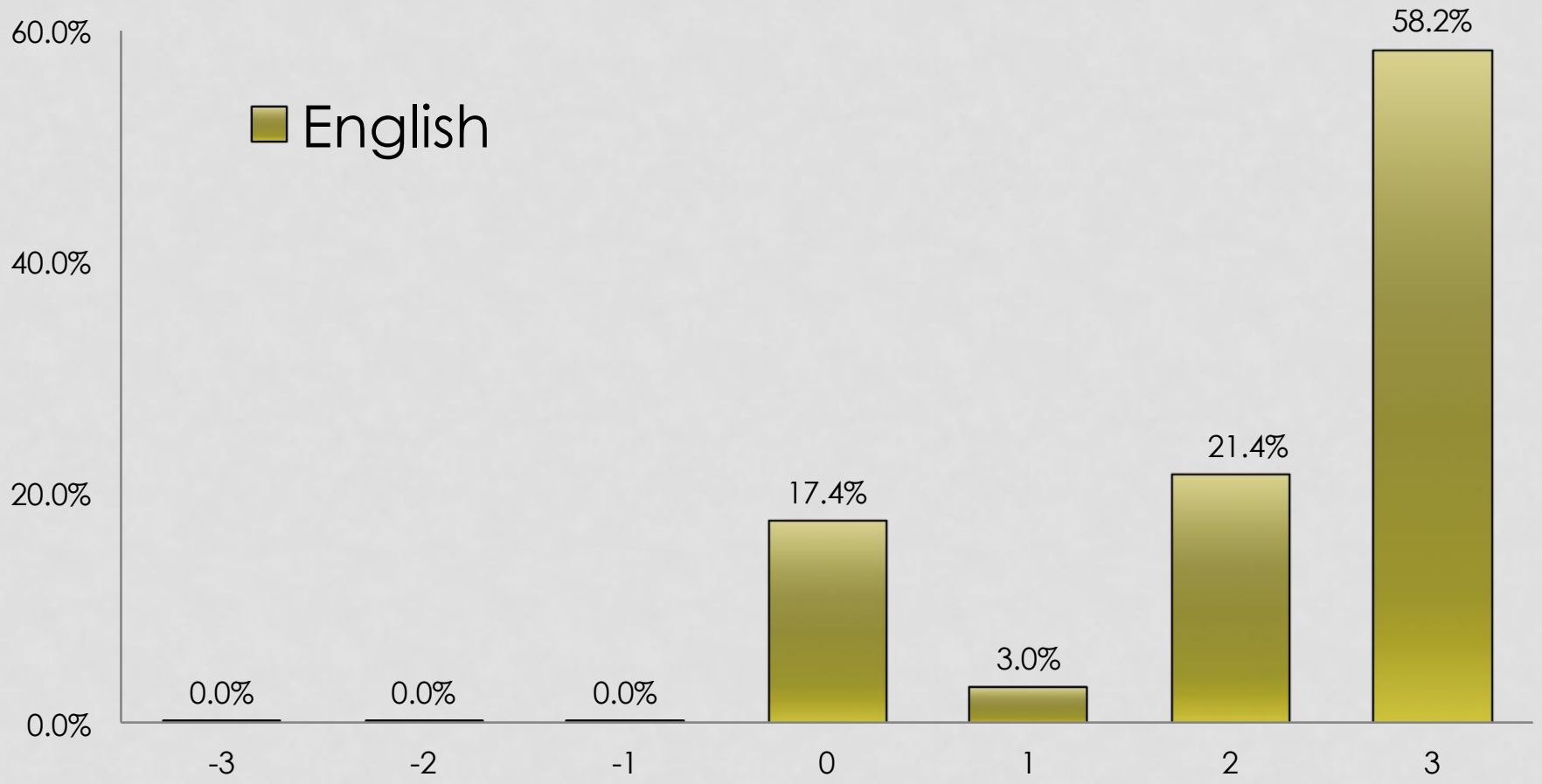
- Berent et al. (2007): Poverty of the Stimulus
  - English speakers exhibit sonority projection effects
    - *[lb]ack (-2) < *[bd]ack (-1) < *[bn]ack (1)
  - Raw lexical statistics don't capture effect
- Daland et al. (2011): No Poverty of the Stimulus
  - Several lexicalist models derive SSP for English
  - These models can capture experimental results
  - They do not need build in SSP preference
  - As long as they have:
    - **Syllable structure** tells the model [gb] in rug.by doesn't count
    - **Features** tell the model what sounds are similar to one another
    - **Frequency Sensitivity** allows models to favor more frequent patterns
      - *#[+son][-son] vs. *#[-son][+son]
      - Captures preference for #[**bn**]ack over #[**nb**]ack
      - More words in English similar to [bn] than to [nb]
- Raw Statistics not sufficient. CBG required.

ENGLISH SYLLABLE ONSETS

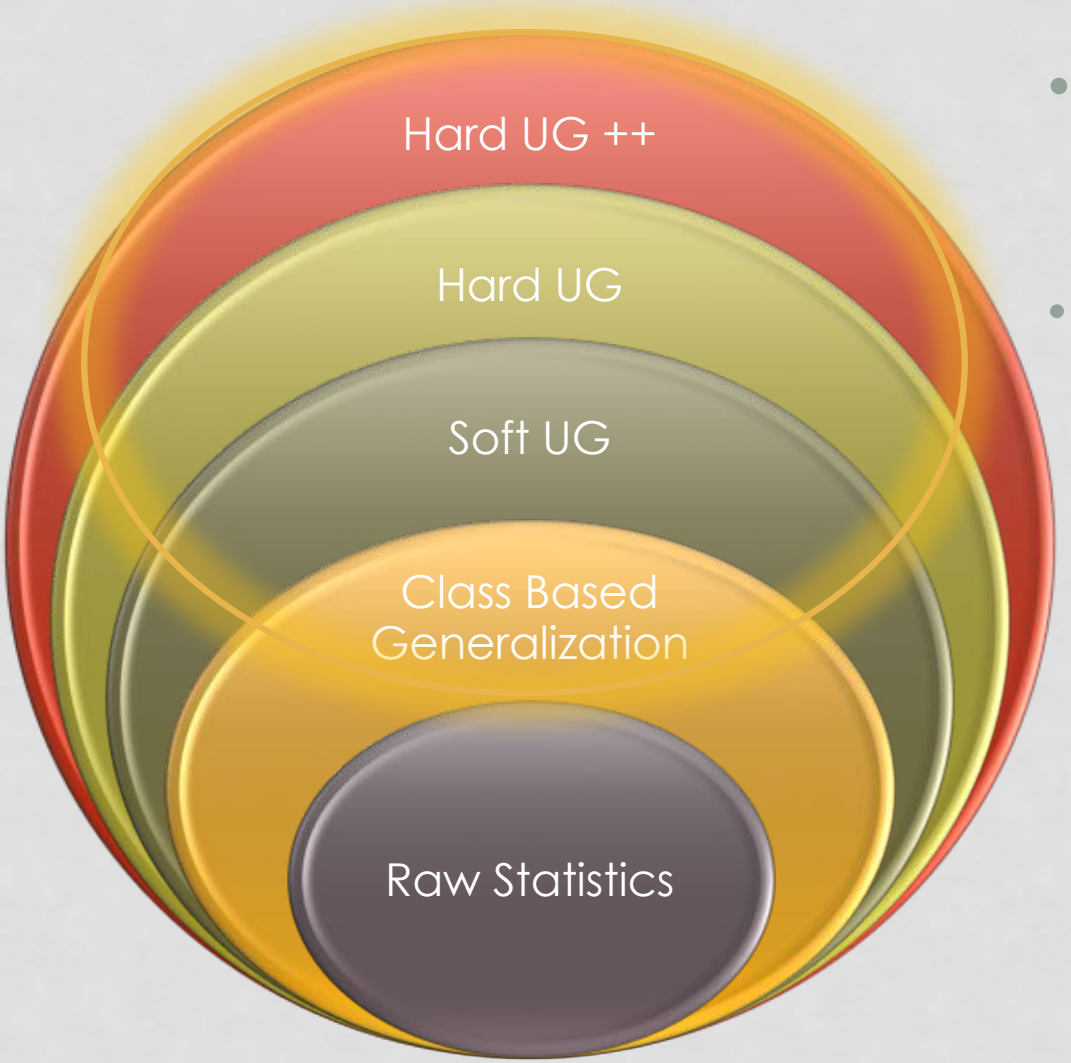
OO (0)		ON (1)		OL (2)		OG (3)	
st	521	sn	109	fl	290	pr	1046
sp	313	sm	82	kl	285	tr	515
sk	278			pl	238	kr	387
				bl	213	gr	331
				sl	213	br	319
				gl	131	fr	254
						dr	211
						kw	201
						sw	153
						hw	111
						θr	73
						tw	55
						ʃr	40
						dw	17
						gw	11
						θw	4
(17.4%)		**(3.0%)**		**(21.4%)**		**(58.2)%**	

(data from Hayes & Wilson 2008)

ENGLISH



CBG IS SUFFICIENT?

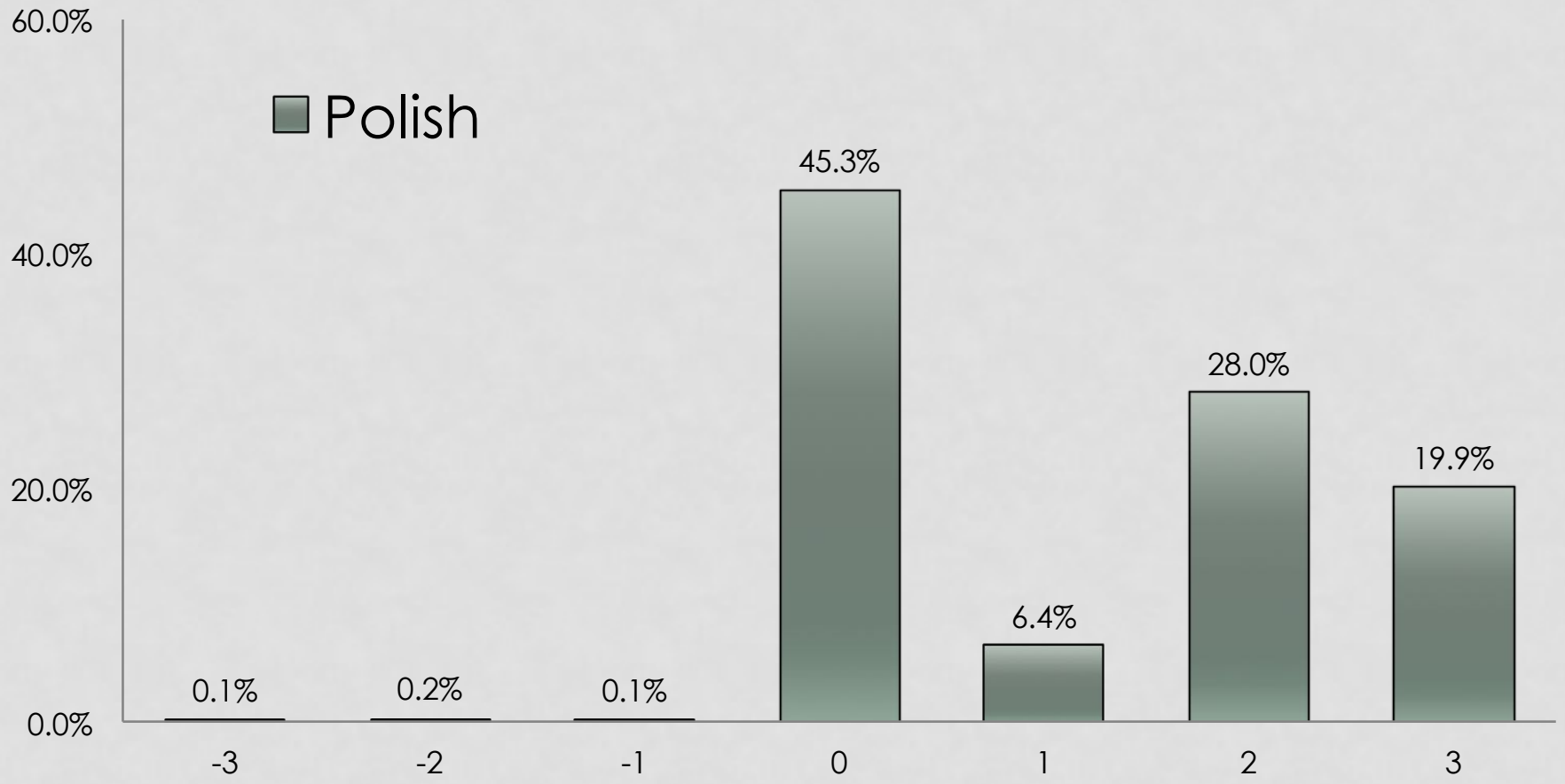


- CBG is enough
  - Daland et al (2011)
  - Hayes (2011)
- Jarosz (under review)
  - English is not the right test case
  - We need language whose input ***contradicts*** SSP statistically

THE OPPOSITE?

- Polish?
  - Whole Scale!
  - [wb]ack < [lb]ack < [mb]ack < [bd]ack < [bn]ack < [bɹ]ack < [bj]ack
  - -3 -2 -1 0 1 2 3
  - [wza] [lvɨ] [mʂa] [ptak] [dnɔ] [klutʃ] [zwɨ]
- Polish Child Directed Speech Sample
  - From Polish CDS Frequency Dictionary (Haman 2011)
    - ~800k word tokens (~115k #CC)
    - ~44k word types (~11k #CC)

THE OPPOSITE: POLISH?

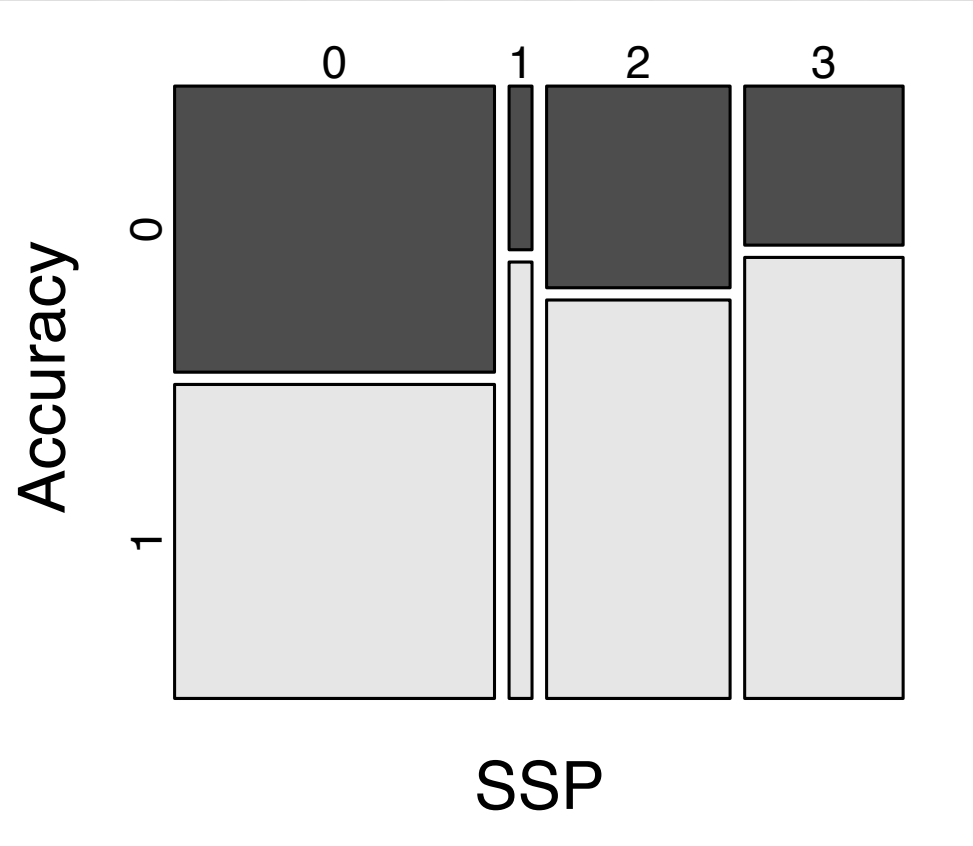


QUESTIONS

- What do Polish speakers know about SSP?
  - Previous Findings (Jarosz 2016 / under review)
    - Development: Corpus study of spontaneous production
  - New Findings (Jarosz & Rysling 2016 / in prep)
    - Adult phonotactic well-formedness: judgment experiment
- Is the SSP principle derivable from the input?
  - Modeling

ACCURACY BY SSP

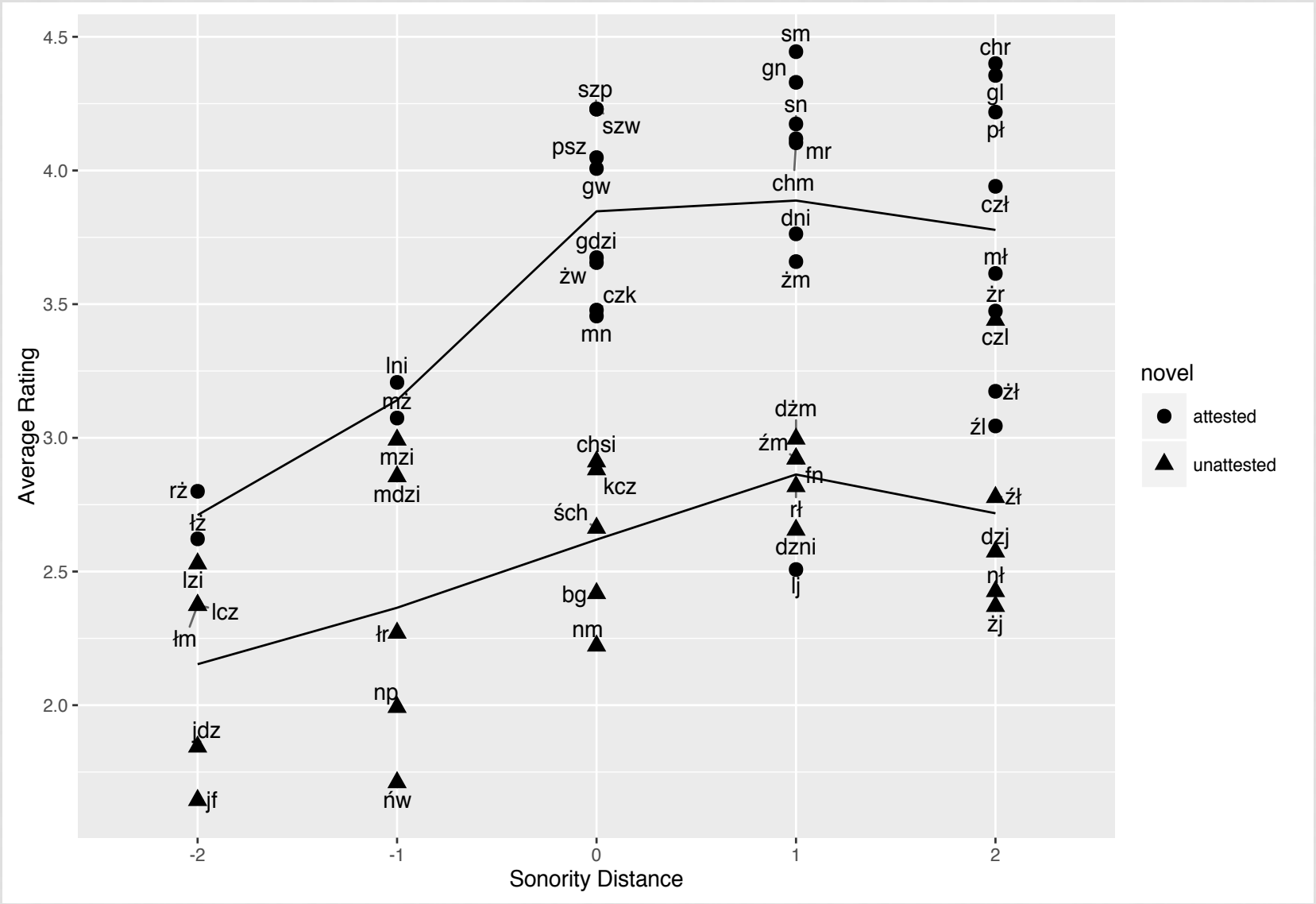
- Spontaneous Production
  - Weist-Jarosz Corpus (Weist et al., 1984; Jarosz 2010; Jarosz et al. 2015)
  - 4 Children (1:7-2;6)
- Raw data visualization
- Just Plateaus & Rises
  - Children didn't attempt falls
- Accuracy rises with SSP
  - **SSP** (***β*=0.28**, *Z*=7.16)
  - Even after controlling for
    - Age
    - Length of target word in syllables
    - Log Word Frequency
    - Primary Stress
    - Participant
    - Function word
    - Morphologically complex



JAROSZ & RYSLING (2016 / IN PREP)

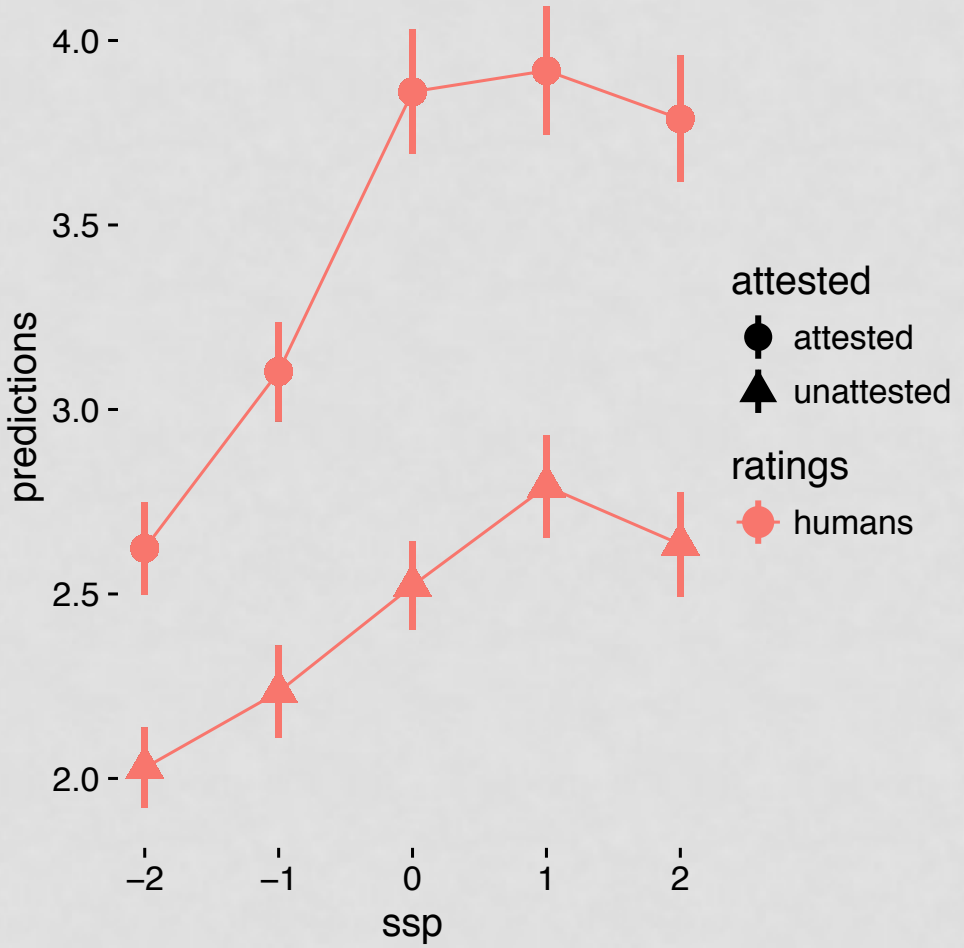
- What happens with adults? In judgments? Across the entire scale? Are attested and unattested clusters different?
- Stimuli
  - **28 attested heads** ranging in Sonority Rise -2/-3 thru +2/+3
  - **25 unattested heads** ranging in Sonority Rise -2/-3 thru +2/+3
  - **30 tails** ranging in morphological category
  - **10 counter-balanced presentation lists**
    - 159 test items (53 heads X 3 tails)
    - 240 fillers introducing variation in length and onset shape
- Procedure
  - Each nonce word presented orthographically
  - Pronounce the word to themselves
  - Rate on how 'natural it sounds' as a Polish word: **scale 1 to 7**
- Participants
  - **81 native Polish speakers**
  - Entirely in Polish, administered online through Polish contacts

RESULTS: AVERAGE RATINGS BY CLUSTER & ATTESTEDNESS



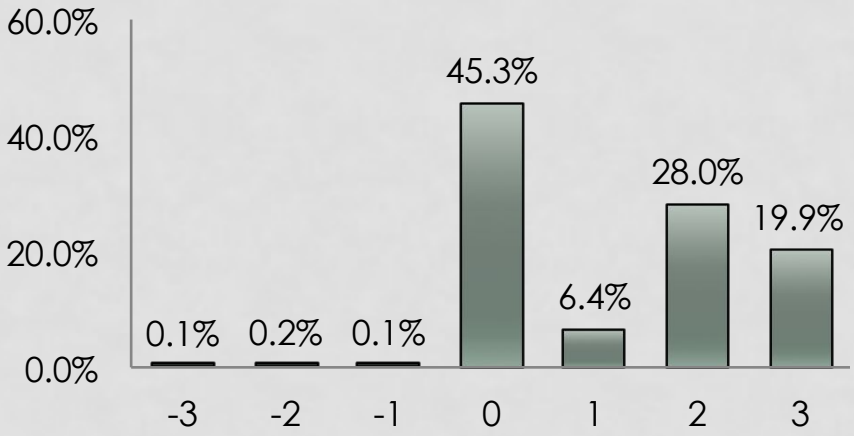
RESULTS

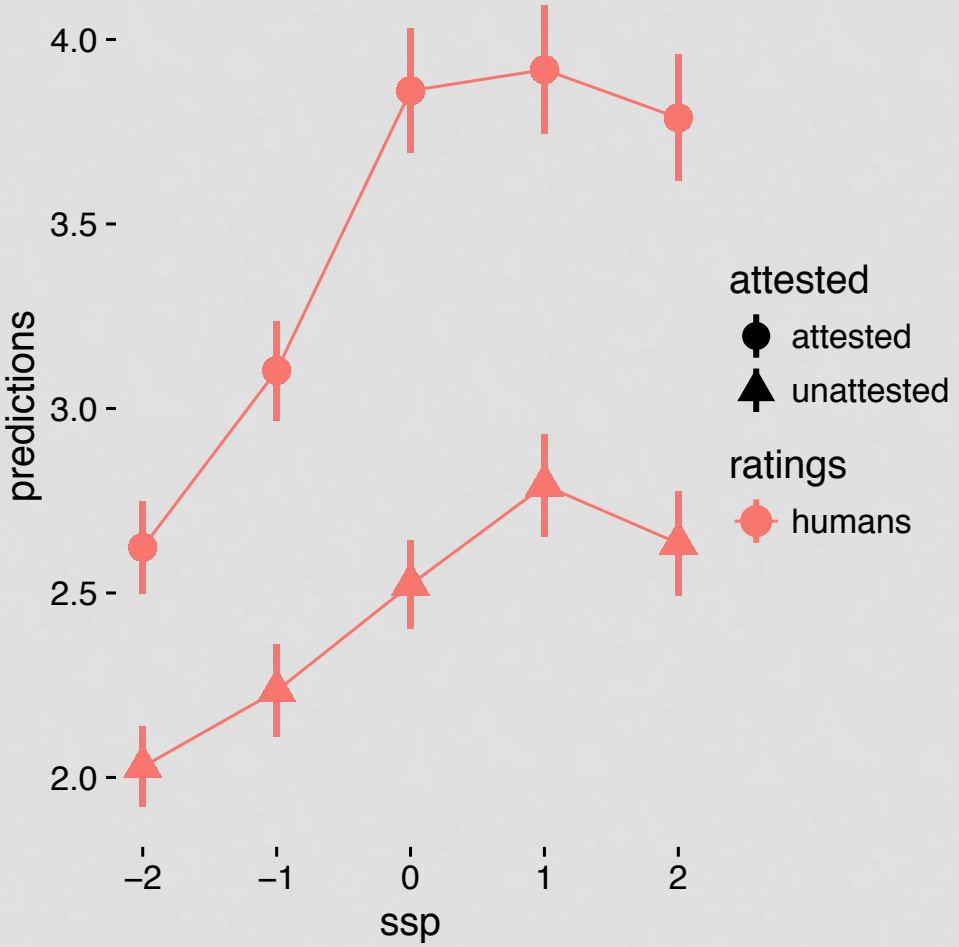
- Mixed effects model with full random effects structure
- Dependent
  - Rating
- Fixed effects
  - SSP * Attestedness
- Random slopes and intercepts, by
  - Subject
  - Tail
- Results
  - **SSP** (β=0.20, t=8.80)
  - **Attestedness** (β=0.57, t=16.18)
  - no significant interaction (β=0.02, t=1.42)



FURTHER DIRECTIONS

- Plateauing?
  - Adults seem to plateau
    - No sig. SSP in 0-3 range
  - Kids rising preferences 0->3
- Replication on 0,1,2s
  - If all true, suggests balance between prior bias and experience
- Remember input skew:

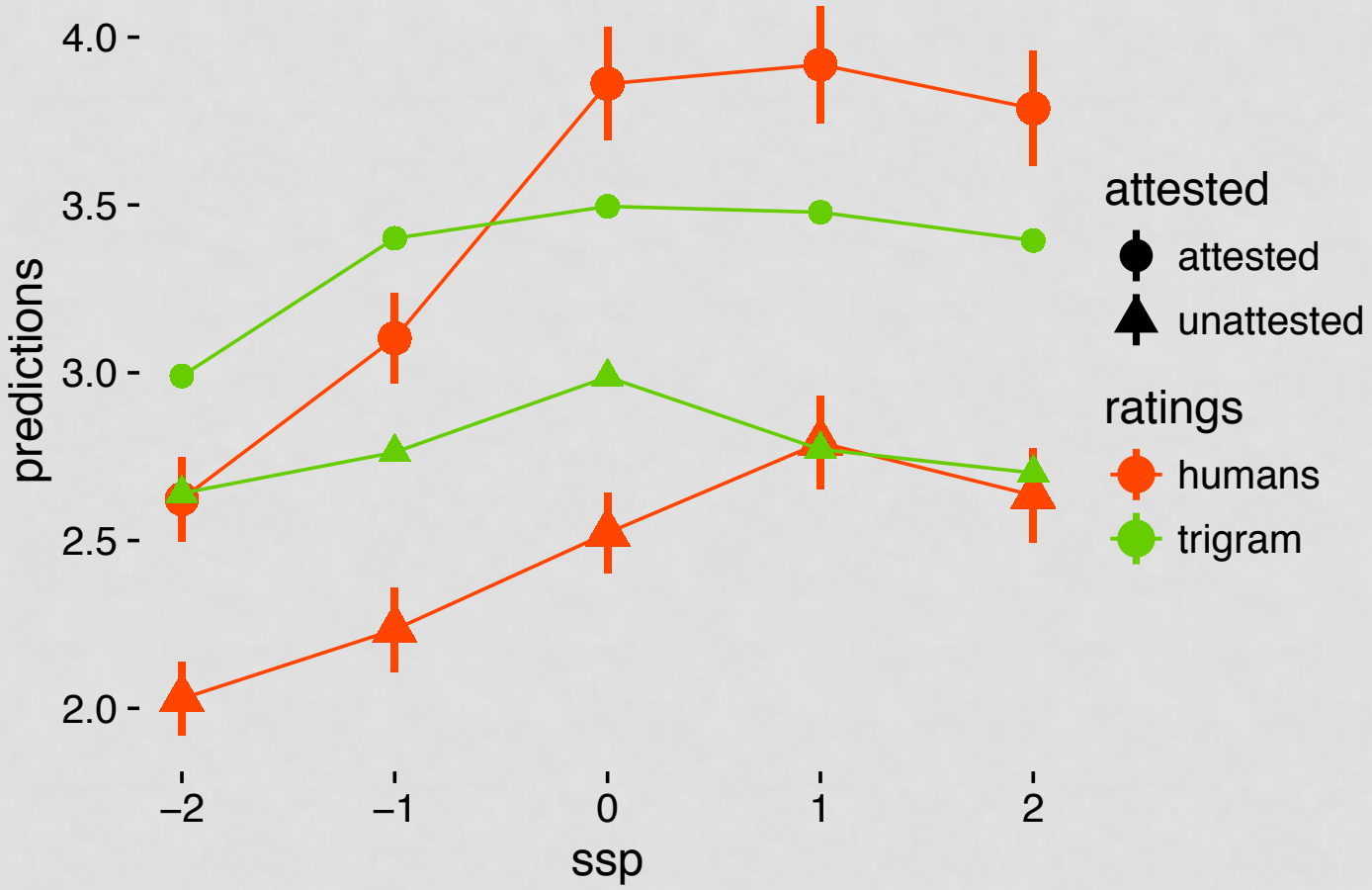




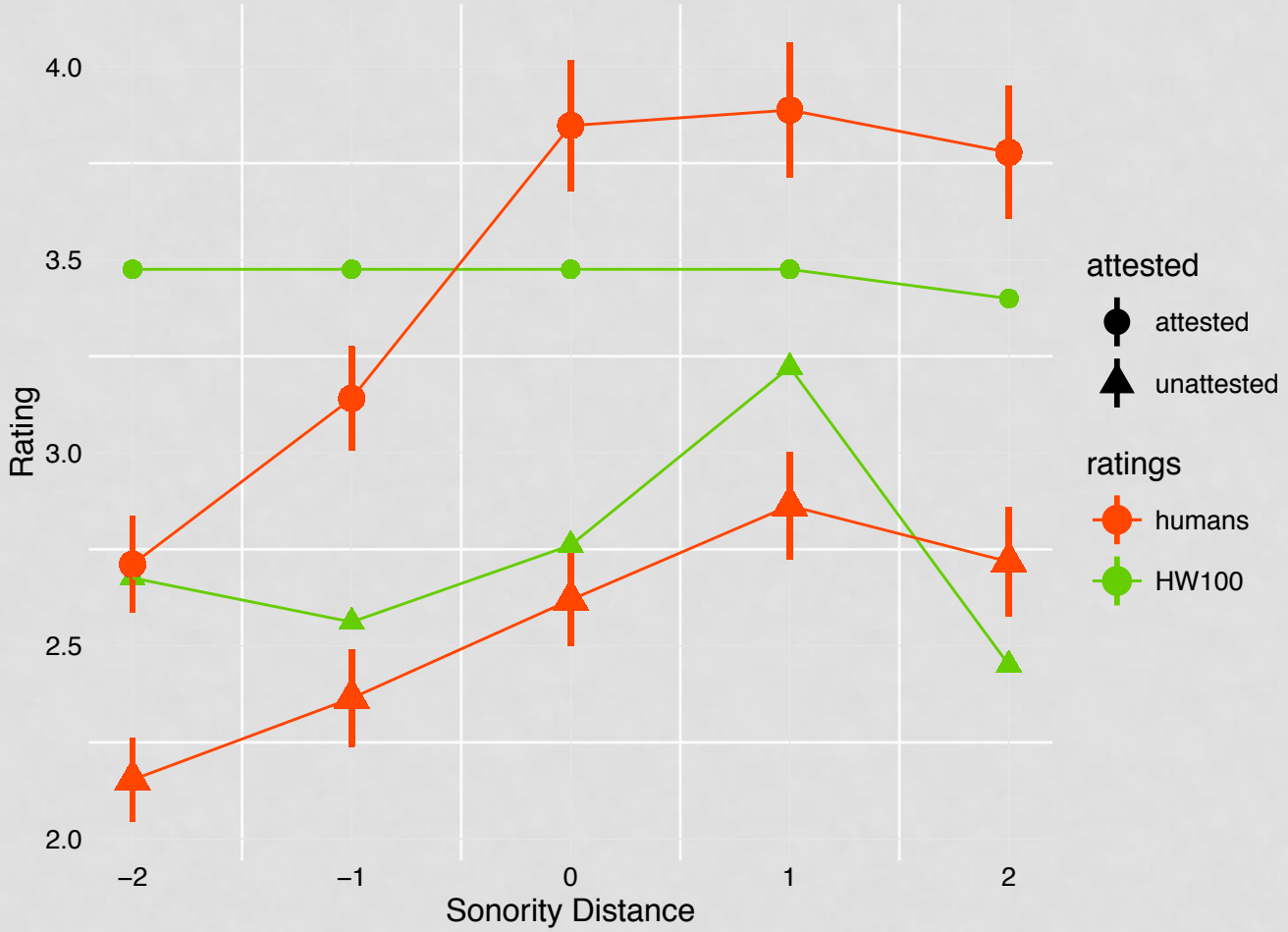
MODELING

- Can models derive the SSP effect from the input?
- Models & Training
  - Trained on phonetically transcribed Polish lexicon (43230 words)
    - Derived from child directed speech to 1;6-3;2
  - Smoothed Trigram
    - As a baseline
  - Daland et al. (2011)
    - Word transcriptions
    - Syllabified word transcriptions
      - Maximal onset with observed word-initial clusters
      - Represented with +/- rhyme feature
    - Induce 100, 200,… constraints
  - Hayes (2011)
    - UG with 32 sonority-regulating constraints
    - No syllabification: constraints don't refer to syllable position
- Evaluation
  - Qualitative
    - Linear regression: fit each model's predictions optimally to ratings
  - Quantitative
    - Correlation ratings & scores (Overall, attesteds, unattesteds)

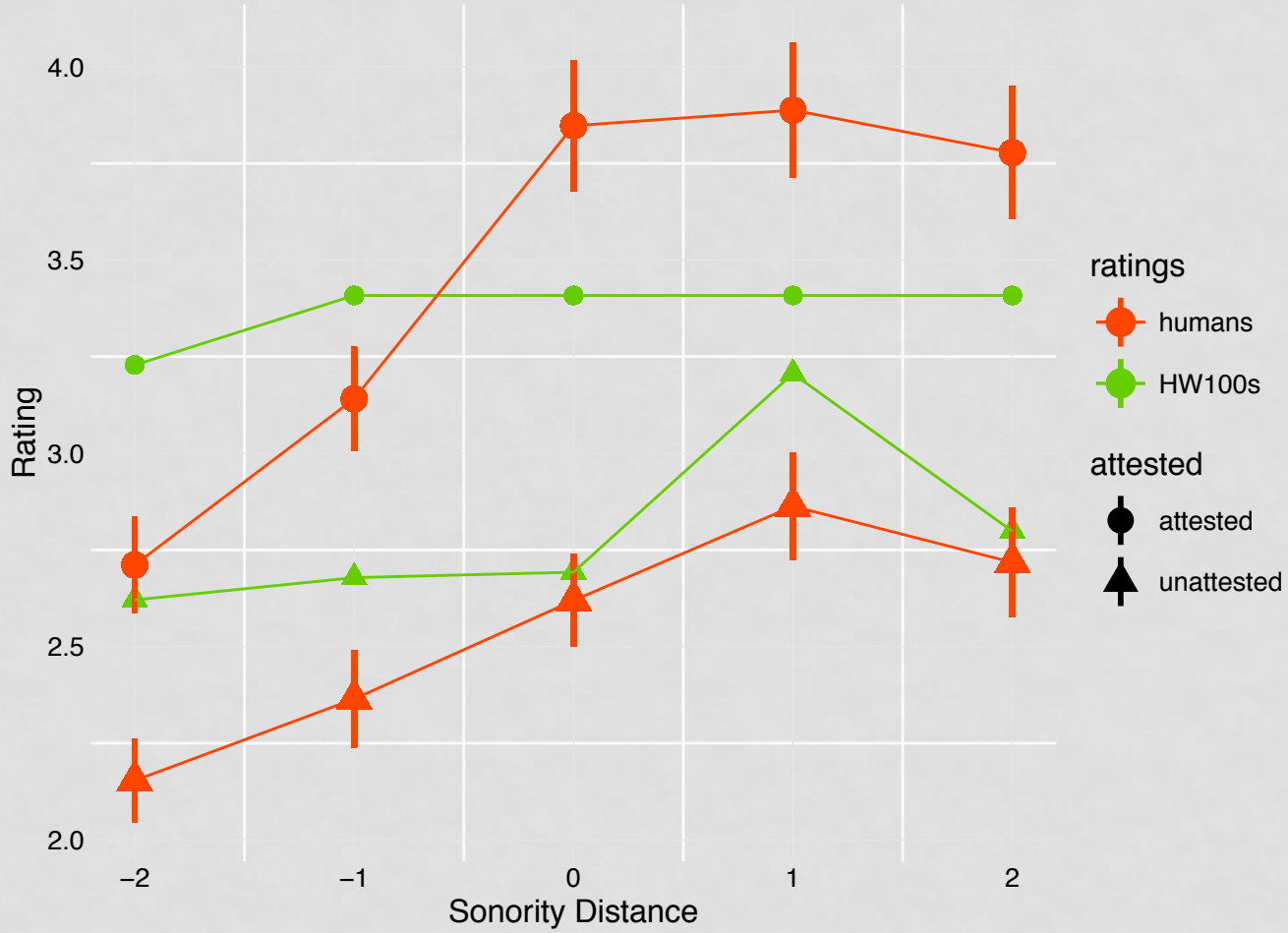
TRIGRAM V. HUMANS



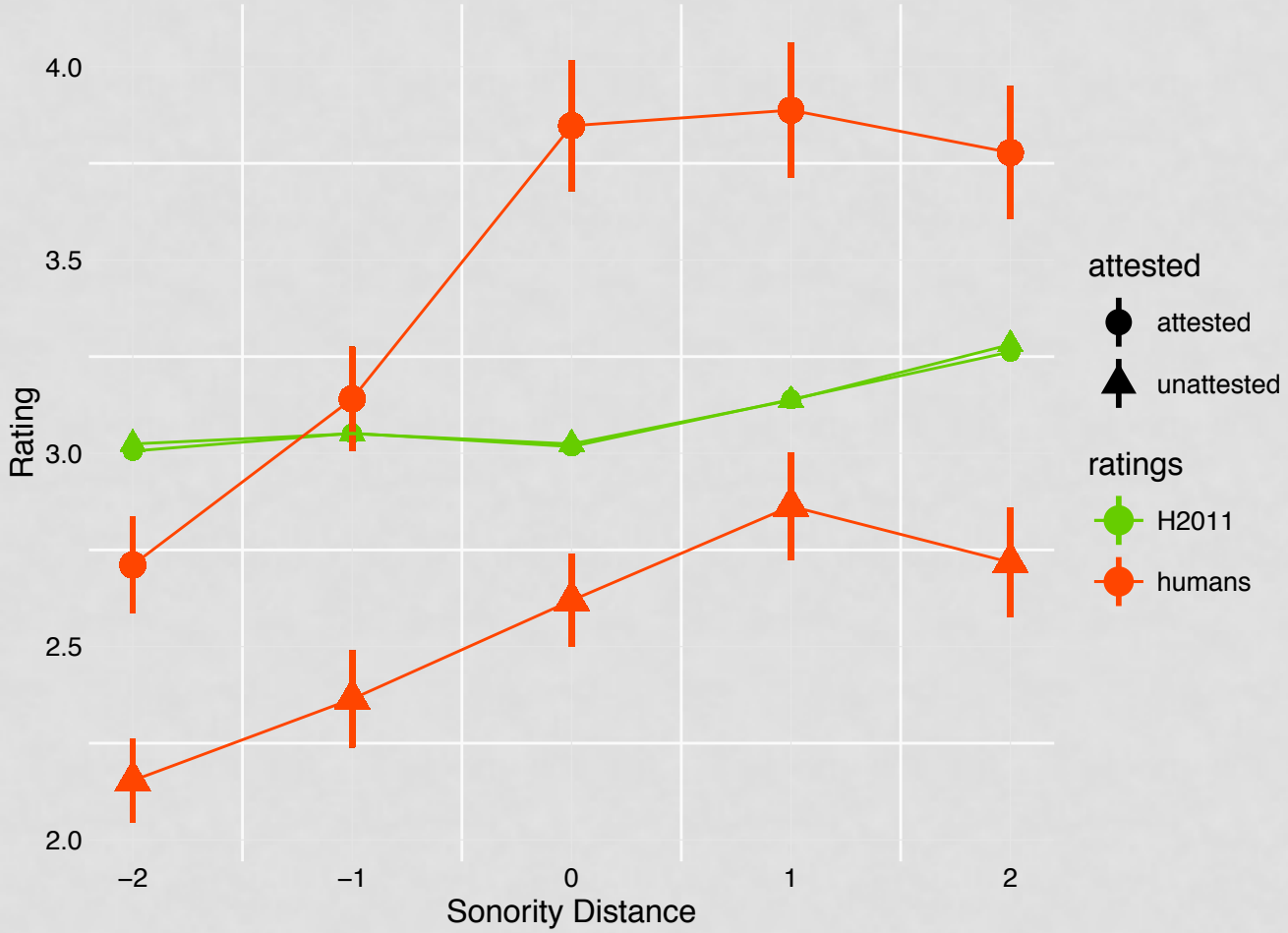
HW2008 V. HUMANS



HW2008 V. HUMANS



HAYES 2011 V. HUMANS

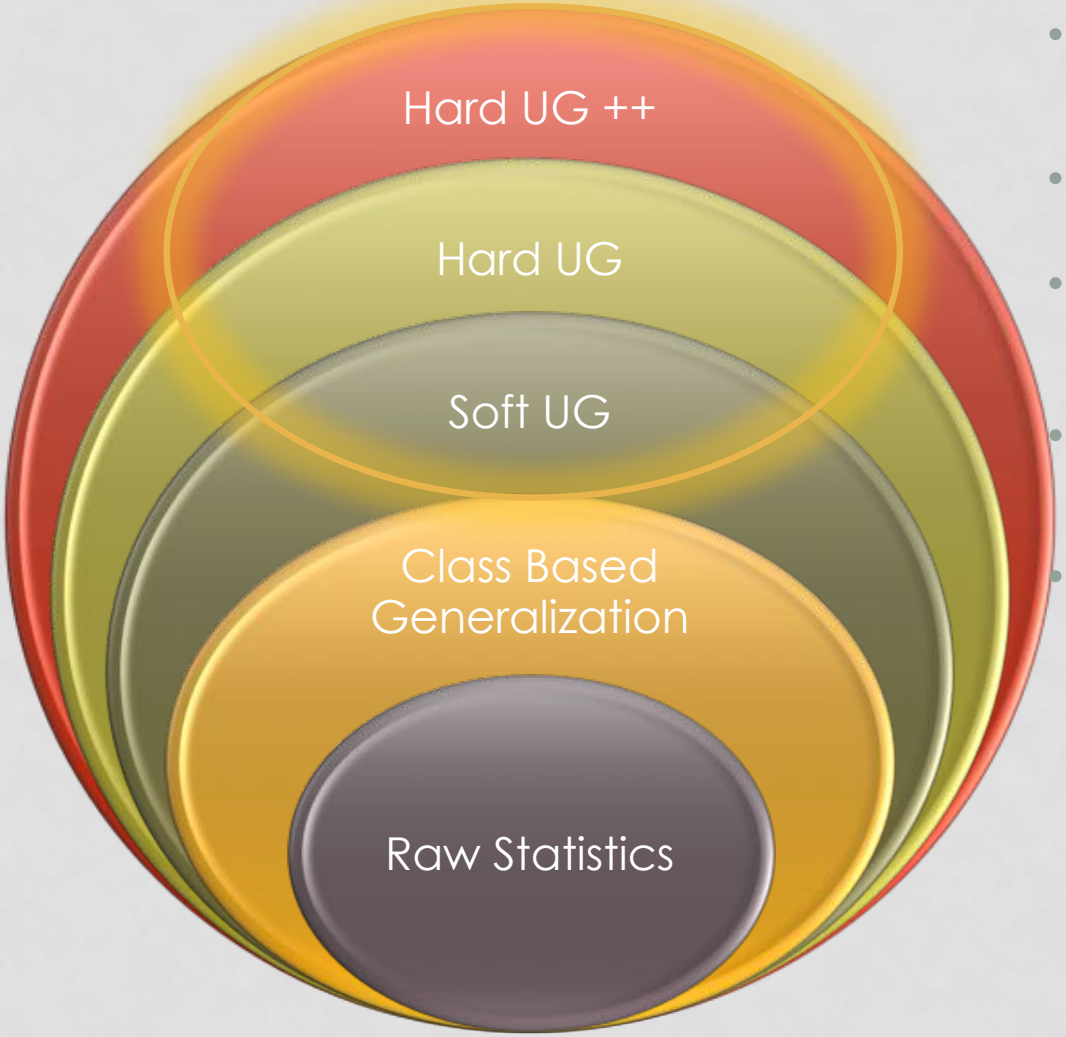


QUANTITATIVE EVALUATION

model	Overall r	Attested r	Unattested r
Trigram	74.2	51.8	25.3
HW2008 unsyll	64.1	8.2	44.2
HW2008 syll	59.4	38.0	39.0
H2011	13.4	1.0	23.5

- These models do not capture SSP
  - Overall correlations for HW2008 version are ok
  - H2011 UG fares poorly across the board
- Unconstrained generalization from phonetic transcriptions of Polish words does not give rise to SSP

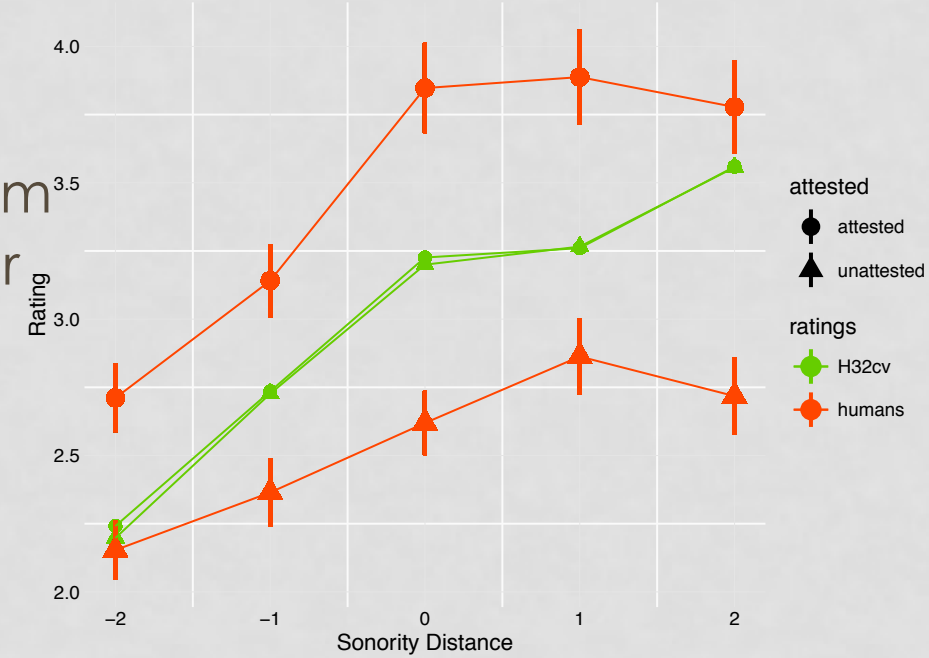
POLISH SSP: DISCUSSION



- Development
  - Children favor clusters with higher rises
- Judgments
  - Adults favor higher rises too
- Just a small sample of models we tested
  - No CBG model succeeded
- Conclusion
  - CBG models cannot be the full story
- Growing literature
  - Becker et al. (2011, 2012), Hayes et al. (2009), Hayes & White (2013), Garcia (2014, 2016)

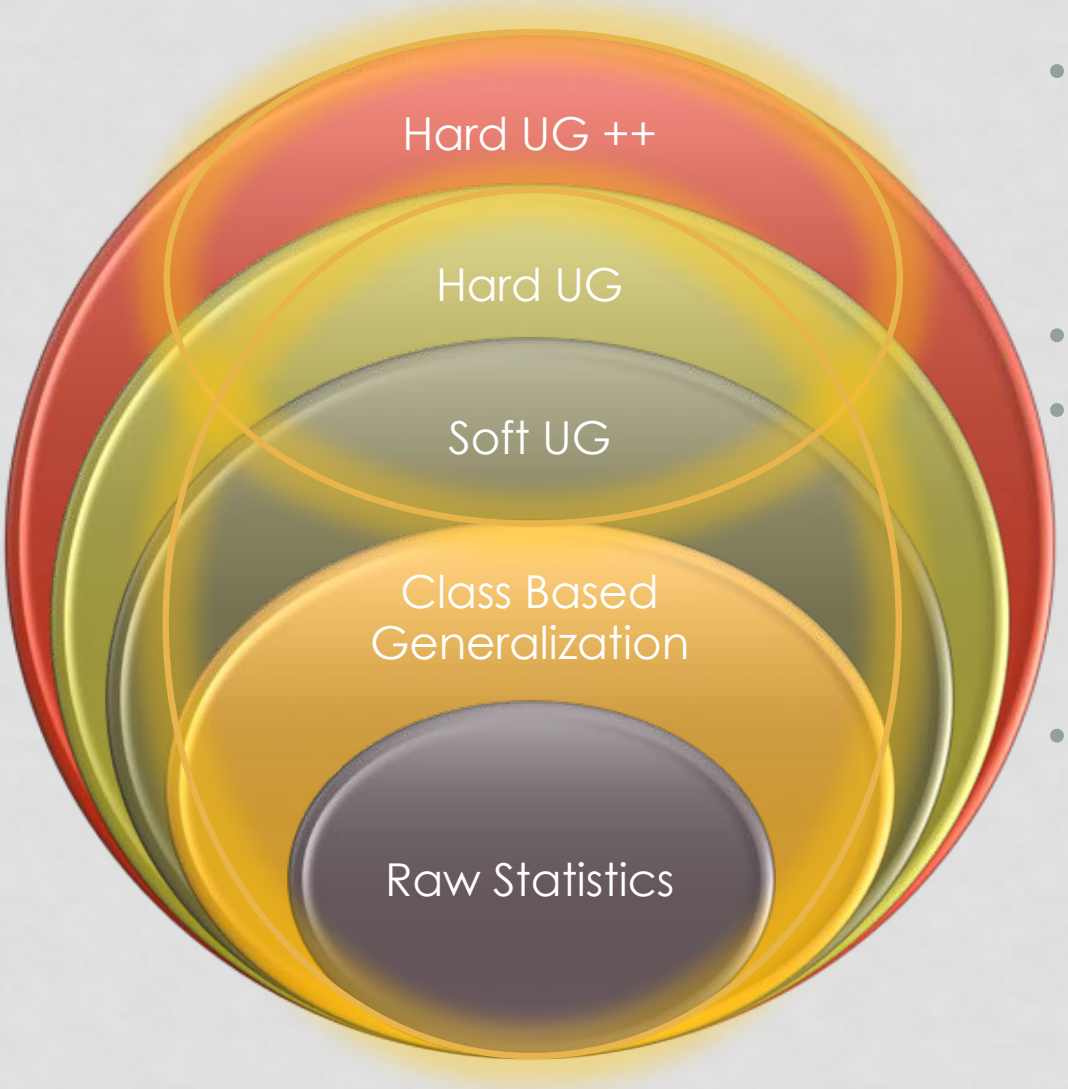
ONGOING: WHAT CAN DERIVE SSP?

- Hard UG is appealing
  - Stringency constraints?
  - But, what prevents learner from undoing this as they learn other phonotactic constraints?



- Constraints on generalization by syllable context?
  - Stand-in proof of concept: only from #C_0V?
  - But, what about phonotactics that are independent of syllable structure?

FINAL COMMENTS



- Case Studies
  - Process Interactions
  - P&P Stress
  - SSP & Phonotactics
- Still much more to do!
- UG + statistical learning interactions
  - Important
  - Interesting
  - Surprising
- Modeling interactions
  - Identify & refine more nuanced hypotheses about how nature and nurture interact