Computation, learning, and typology

Class 5: Data shapes learning

Adam Albright and Eric Baković

CreteLing 2023 — July 2023

creteling2023.phonology.party

Mid-point pathologies

Picking up from last time

- Biases introduced by the learning algorithm
- Learning trajectory: some grammars are ‘closer to the initial state’ than others
  - Fewer steps
  - Higher prior probability
- If data is completely ambiguous, this creates a bias for certain analyses over others
  - E.g., antepenultimate stress $>$ midpoint stress

Extending this result: Stanton (2016)

- So far, we've seen why learners who receive ambiguous data won't arrive at a 'mid-point' analysis
- The problem: midpoint systems are (basically) unattested altogether
- Traditional "UG solution": formulate constraints so that mid-point systems can't be represented
  - And, therefore, can't be learned
- Alternative: data that distinguishes them from antepenultimate stress is rare enough that the bias for antepenultimate stress can exert itself

Stanton's observation

/σσ/	*ExtLapse(L)	*ExtLapse(R)	Align(L)	Align(R)
☞ a. όσ				*
b. σό			*! W	L

/σσσ/	*ExtLapse(L)	*ExtLapse(R)	Align(L)	Align(R)
☞ a. όσσ				**
b. σόσ			*! W	* L
c. σσό			*!* W	L

/σσσσ/	*ExtLapse(L)	*ExtLapse(R)	Align(L)	Align(R)
a. όσσσ		*! W	L	*** W
☞ b. σόσσ			*	**
c. σσόσ			**! W	* L
d. σσσό	*! W		*** W	L

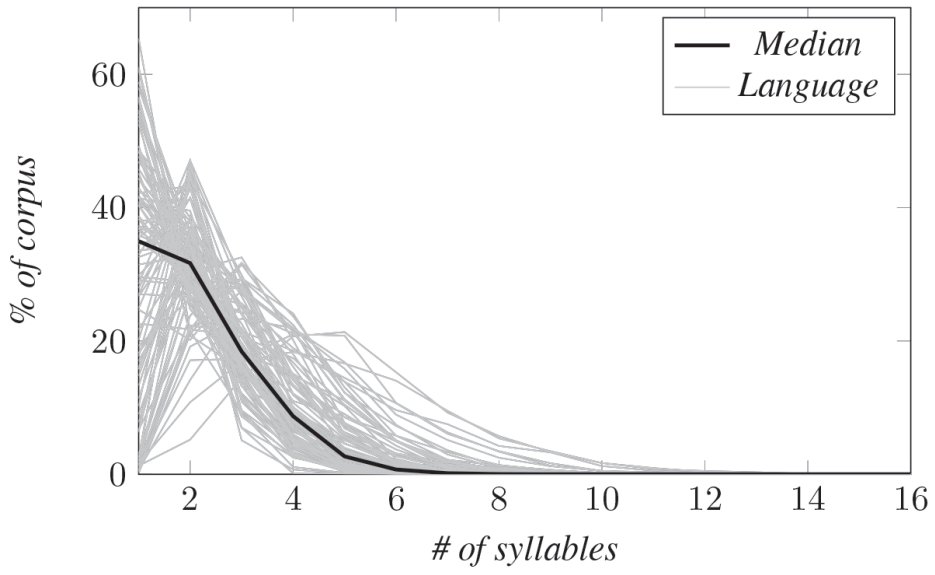
/σσσσσ/	*ExtLapse(L)	*ExtLapse(R)	Align(L)	Align(R)
a. όσσσσ		*! W	L	**** W
b. σόσσσ		*! W	* L	*** W
☞ c. σσόσσ			**	**
d. σσσόσ	*! W		*** W	* L
e. σσσσό	*! W		**** W	L

/σσσσσσ/	*ExtLapse(L)	*ExtLapse(R)	Align(L)	Align(R)
☞ a. όσσσσσ		*		*****
b. σόσσσσ		*	* W	**** L
c. σσόσσσ		*	** W	*** L
d. σσσόσσ	*! W	L	*** W	** L
e. σσσσόσ	*! W	L	**** W	* L
f. σσσσσό	*! W	L	***** W	L

/σσσσσσσ/	*ExtLapse(L)	*ExtLapse(R)	Align(L)	Align(R)
☞ a. όσσσσσσ		*		******
b. σόσσσσσ		*	* W	***** L
c. σσόσσσσ		*	** W	**** L
d. σσσόσσσ	*! W	*	*** W	*** L
e. σσσσόσσ	*! W	L	**** W	** L
f. σσσσσόσ	*! W	L	***** W	* L

- Clear evidence for Align(L) ≫ Align(R) in 2,3,4-syllable words
- Evidence for *ExtLapse(R) ≫ Align(L) from 5-syllable words
- Evidence for *ExtLapse(L) ≫ *ExtLapse(R) only from 6-syllable words and longer

On the relative scarcity of long words



#	Median
1	35%
2	32%
3	18%
4	9%
5	3%
6	1%
7	0%

FIGURE 1. Results of the survey of text corpora from 102 languages (see the appendices for more details).

- Rough estimate of relative proportion of words of different lengths in texts of 102 languages
- With a few notable exceptions, $\geq 6\sigma$ words are a very small proportion of the input
- Also: long words tend to be morphologically complex (may show other patterns)

Learning especially from short words

- For reasons that we've already discussed, learners exposed to just the $< 6\sigma$ data of midpoint system would infer antepenultimate stress
- Even with exposure to 6+σ words, majority of updates are for the "antepenultimate stress" subset of the rankings
- Possible result: non-zero probability that midpoint stress mislearned as antepenultimate stress
- Question: is distribution of the data skewed enough to impact typology, when transmitted across generations?

Iterated learning

Hypothesis

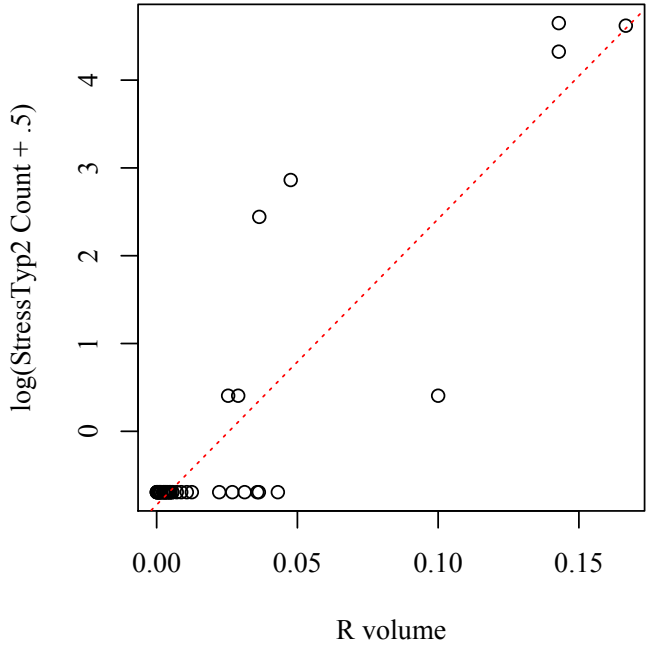
- Typological frequency does not reflect r-volume directly, but the learned r-volume
  - Distribution over rankings that emerge as a result of learning across generations
- Albright and Subramaniam (2109): tested with an iterated learning model

Baseline: r-volume

- Along with Bane & Riggle (2008), Riggle (2010), Staubs (2014), Stanton (2016) and others, we assume a baseline in which grammars are sampled randomly
- R-volume
- Calculated r-volume for each of the 48 predicted patterns, using recursive technique proposed by Riggle (2010)

Comparison to empirical data

- As in Bane & Riggle (2008), compared r-volume with count in StressTyp2
- Specifically, log of r-volume



Assessment

- Correlation between (log) r-volume and frequencies of 9 attested patterns: Spearman's $\rho = .612$
- Qualitatively imperfect
  - 39 patterns with non-zero r-volume are empirically unattested
  - Although some of these may be accidental gaps, many are high enough to yield non-zero predicted counts in a database this size (Stanton, 2016)
- Quantitatively imperfect
  - Correlation artificially inflated by large number of unattested patterns
  - Relative frequencies of attested patterns are not well predicted

Hypothesis: iterated learning could improve both aspects of the model

- Difficult to learn patterns are reanalyzed and become rarer or even unattested over time
- Probability mass reassigned, changing relative predicted relative frequency of attested patterns

Learning-Based Frequencies

In order to test this hypothesis, we submitted each of the 48 derivable stress patterns to an iterated learning model, to see how probabilities are reassigned over time

The learning model

- Gradual Learning Algorithm (Boersma, 1997), as modified by Magri (2012) to guarantee convergence
- Error-driven learning
  - Model receives input/output pairs: /σσσσ/, [σσόσ]
  - Model uses current grammar to derive predicted output
  - If incorrect, adjust ranking values to favor trained output
- Parameters
  - All markedness constraints start out equally ranked (100)
  - Plasticity: starts at 1, gradually decreases to .001 (learning slows with age; step=.004)
  - Learning trials: 2000

Training

- Trained the learning model on each of the 48 stress patterns
- Inputs: words of 2–8 syllables
- Shorter words presented more often than long words, according to mean frequencies of word lengths given by Stanton (2016)

2σ	32%	6σ	1%
3σ	18%	7σ	.1%
4σ	9%	8σ	.1%
5σ	3%		

- Each pattern run 1000 times, yielding a probability distribution over learning outcomes

Iterated learning

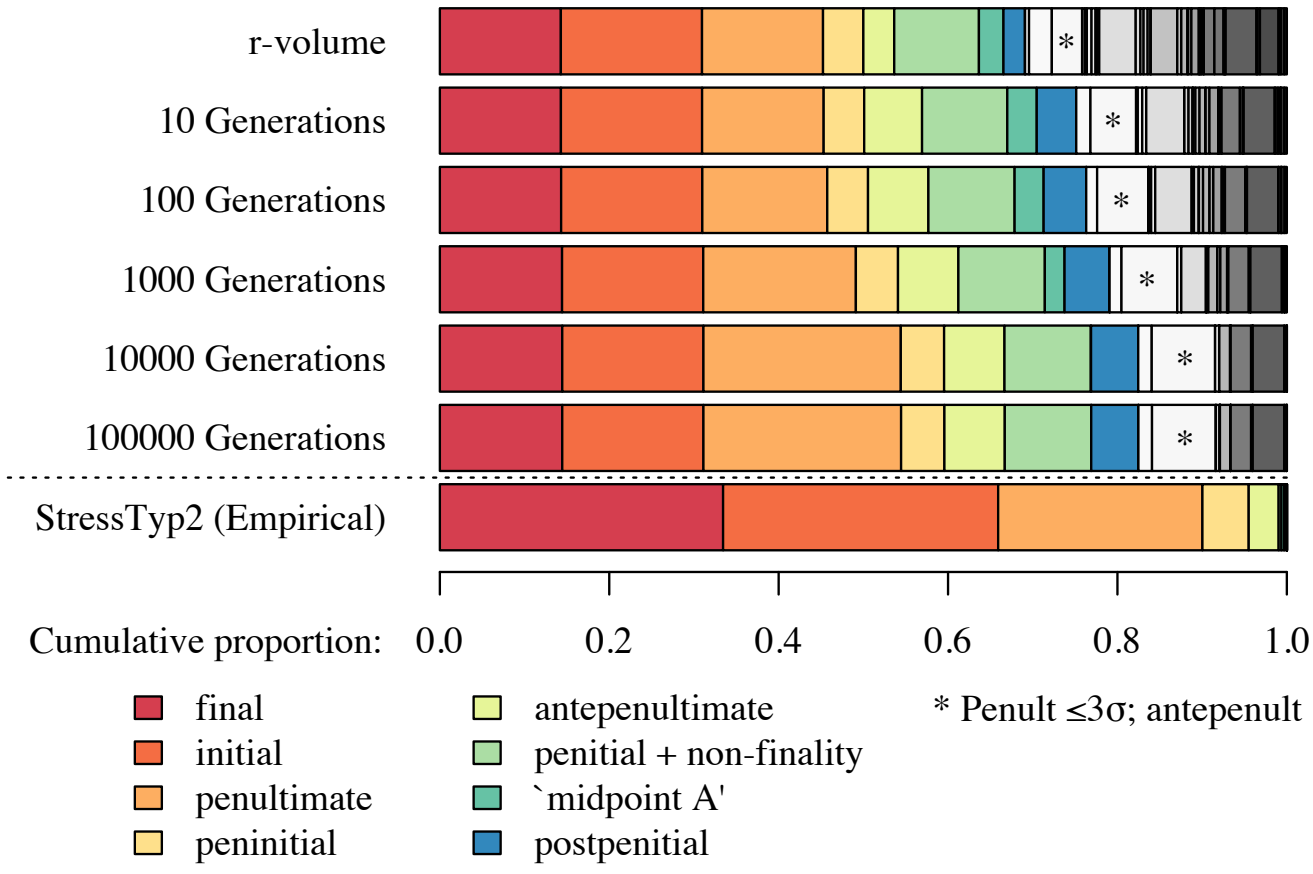
- Starting point: each pattern assigned probability according to baseline distribution (r-volume)
- Imperfect learning changes the probability distribution—e.g., when trained on Pattern 19 (antepenult. $\leq$ 5σ; else peninit.), the model learned:

Pattern 19	(antepenult $\leq$5σ; else peninit)	30.0%
Pattern 20	(antepenult. $\leq$5σ; else postpeninit)	34.1%
Pattern 21	(antepenultimate)	25.4%
Pattern 18	(antepenult. $\leq$5σ; else init)	8.9%
Pattern 23	(antepenult $\leq$4σ; else peninit)	.9%
Patterns 24, 7	(initial, etc.)	$<$1%

- Generations
  - Baseline (r-volume) distribution reassigned according to learned distributions
  - Iterated for 100,000 generations

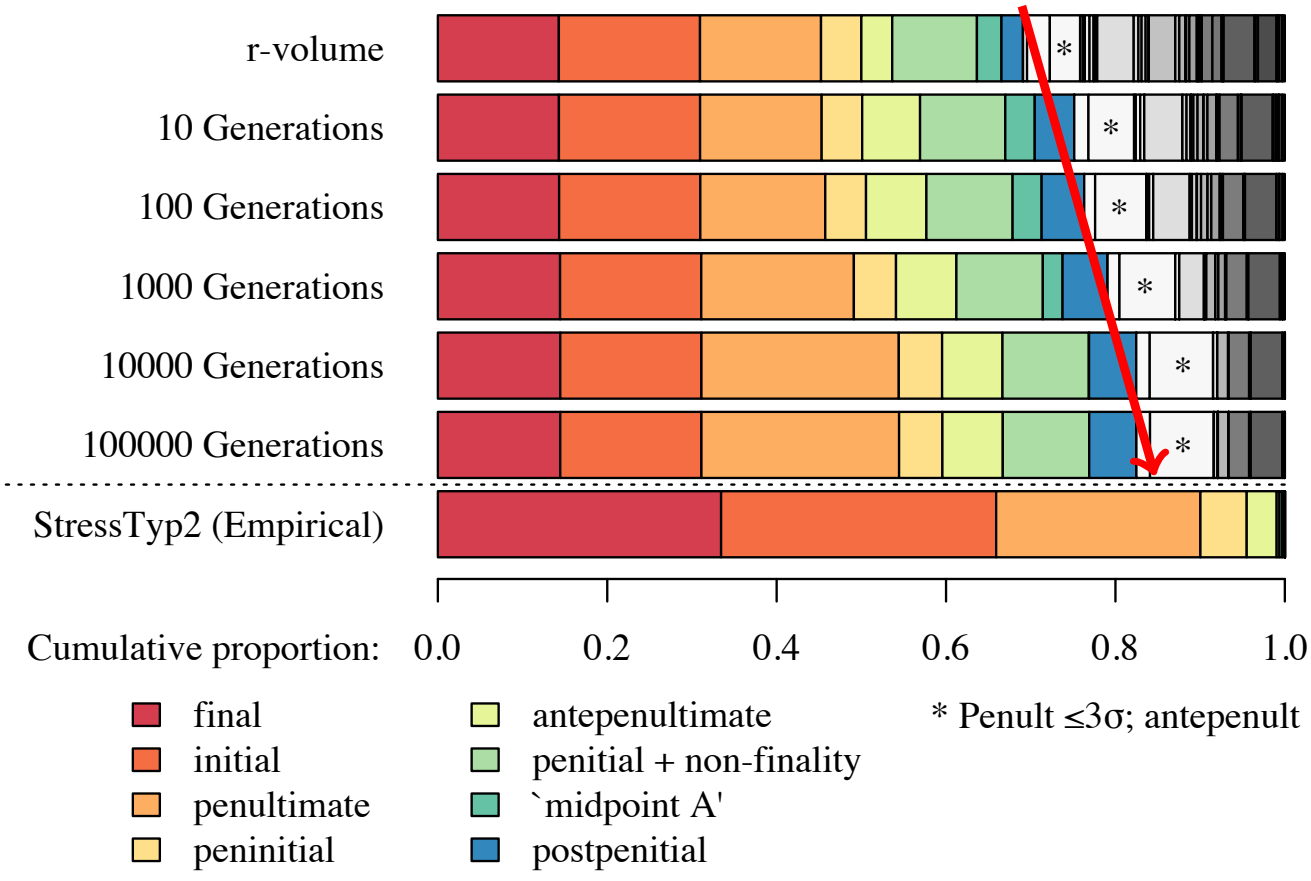
Results

Promising: reduction of unattested patterns



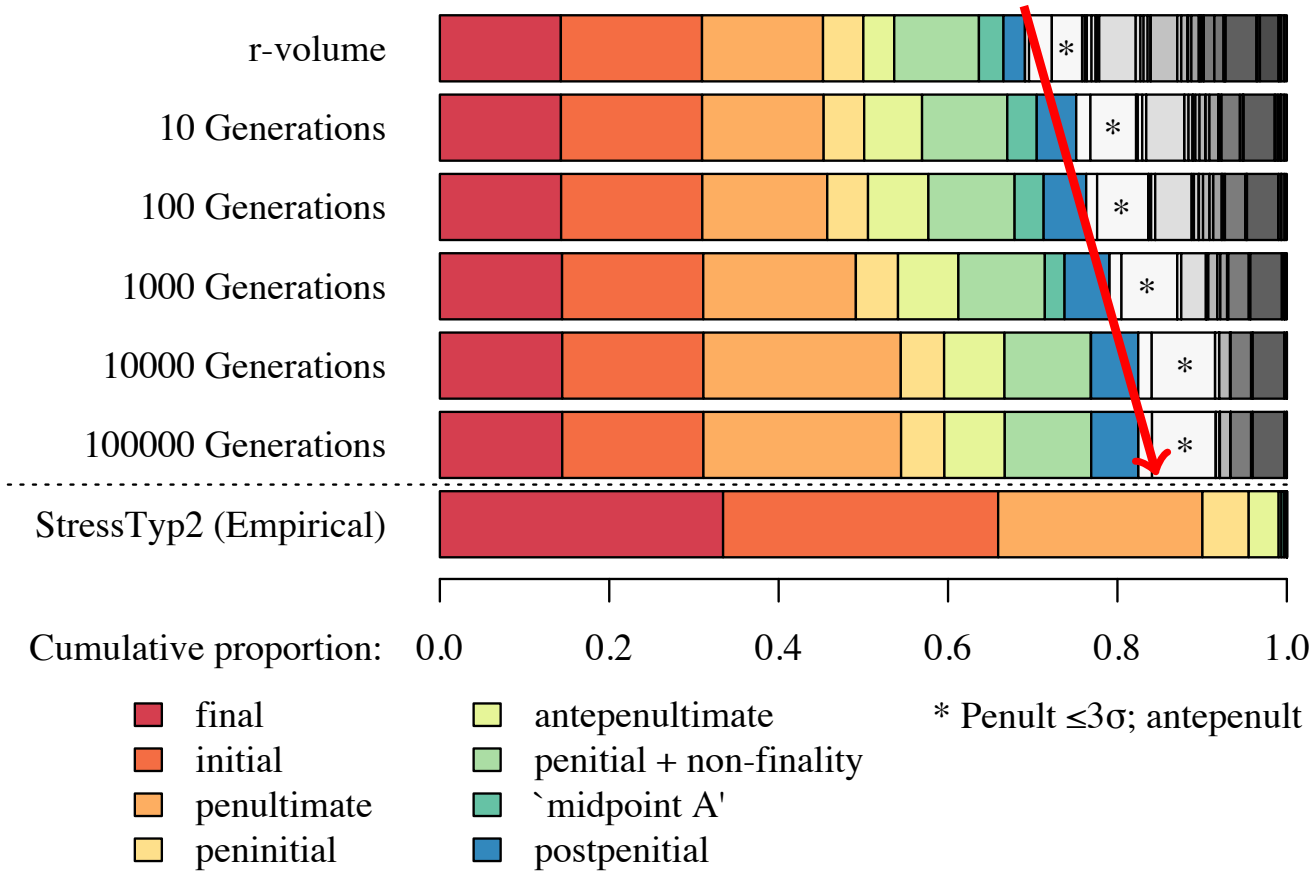
- Patterns attested in StressTyp2 (colored bars) increase in frequency over time

Promising: reduction of unattested patterns



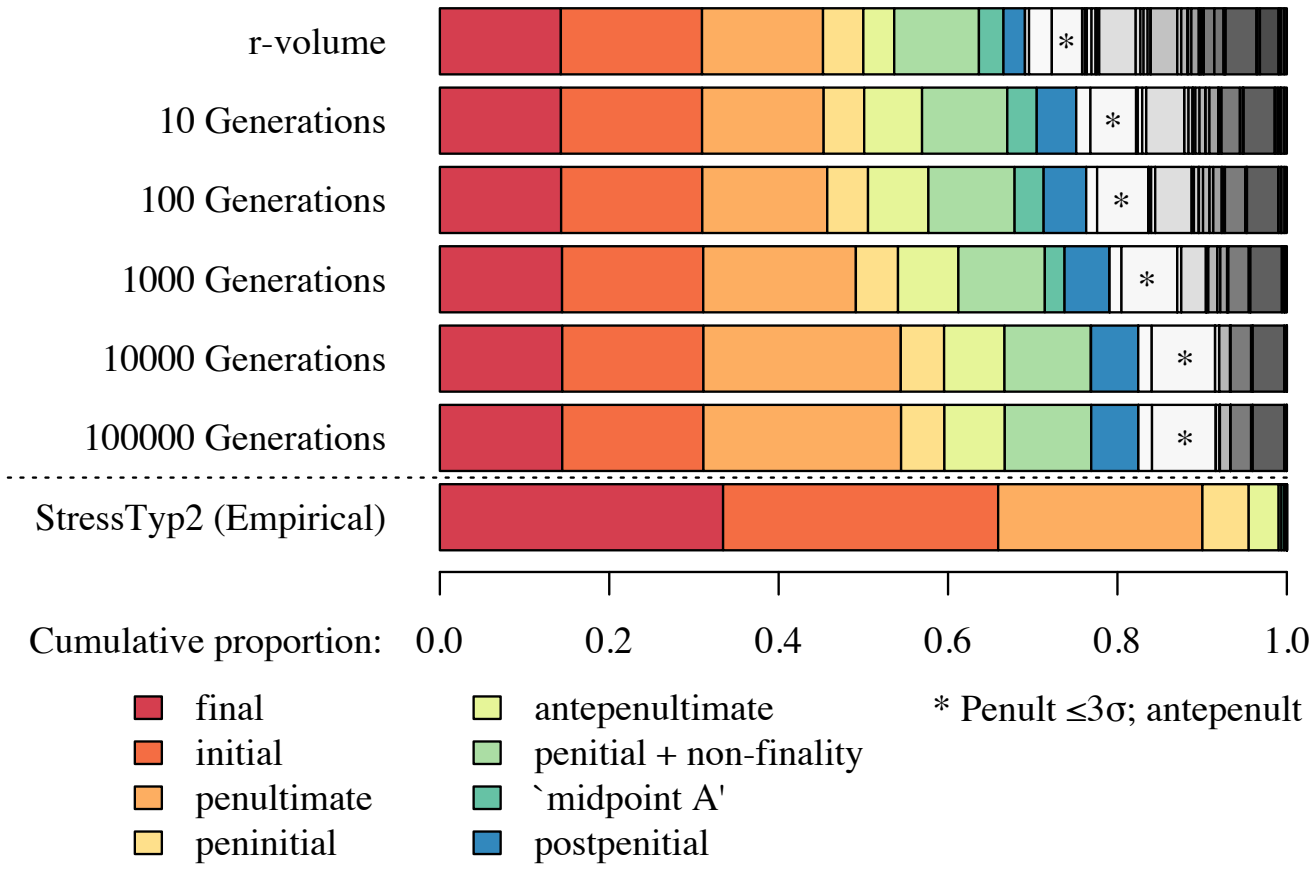
- Patterns attested in StressTyp2 (colored bars) increase in frequency over time

Promising: reduction of unattested patterns



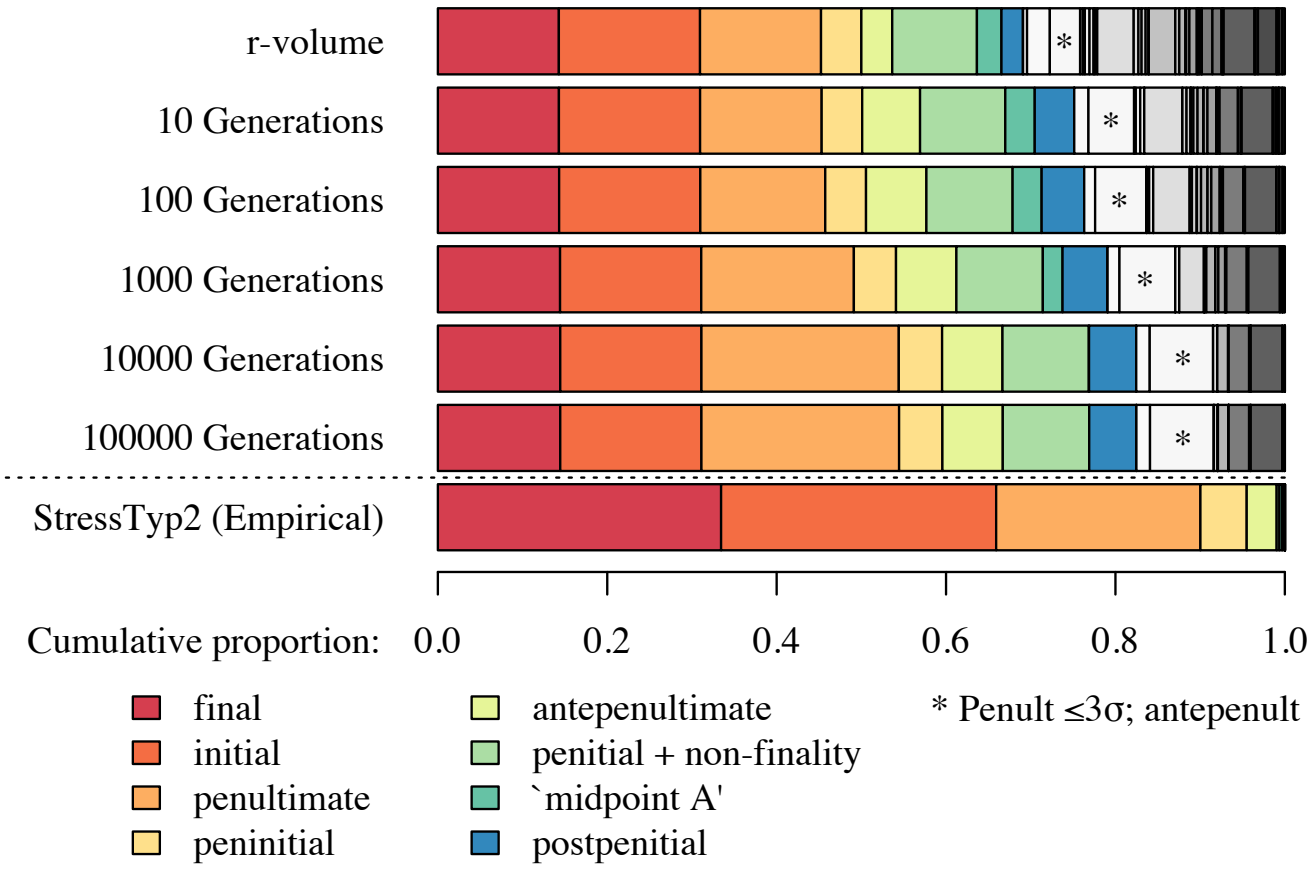
- Unattested patterns (gray bars) decrease in frequency over time

Promising: reduction of unattested patterns



- Some rare or controversially attested patterns also successfully nearly eliminated (‘midpoint A’)

Promising: reduction of unattested patterns



- Consistent with hypothesis that unattested patterns may be eliminated by learning (Blevins, 2004; Staubs, 2014; Stanton, 2016)

Discrepancy 1: stubborn unattested patterns

- Although unattested patterns gradually decrease in relative frequency, some remain or even increase
- Example: '*' penult $\leq 3\sigma$, antepenult longer (a 'midpoint system')

*Lapse, *ExtLapse, *ExtLapse-R, NonFin

$\gg$

Align-L, *Lapse-L, *ExtLapse-L

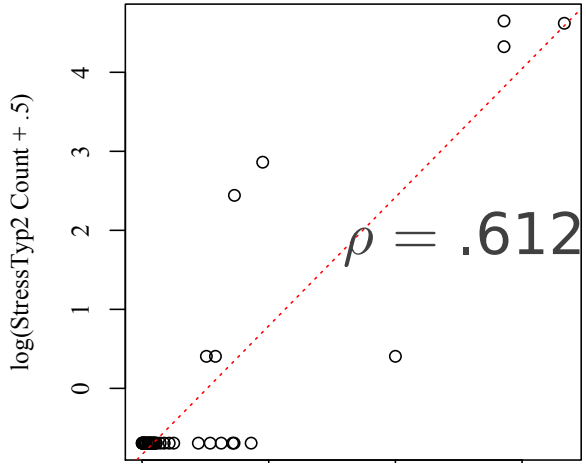
$\gg$

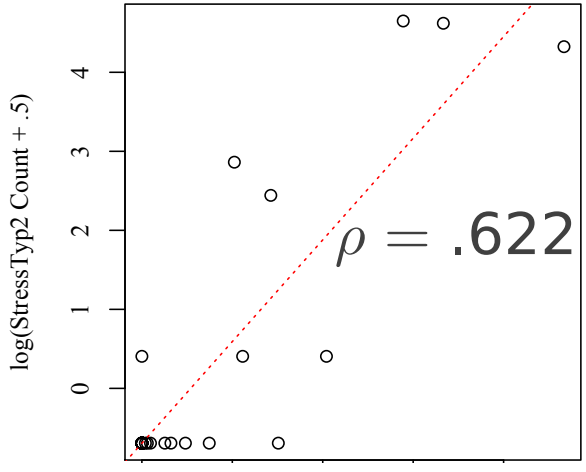
Align-R, *Lapse-R

- This pattern starts out with non-negligible probability (r-volume), and increases as other patterns are reanalyzed as it (does not rely heavily on long words)

Discrepancy 2: unexpected beneficiaries

- Among attested (colored) patterns, predicted final distribution is *less* like attested (bottom row) than in the baseline (r-volume)
- Example: peninitial stress with non-finality ('σσ, σ'σσ)
  - Scarcely attested (Southern Paiute), but easily learned
- Post-peninitial stress is also a popular reanalysis for midpoint systems
- Result: no clear improvement in predictions for relative frequency of attested patterns





Local summary

- An attempt to scale up to a (slightly) more system-wide test of an idea that is intuitive and has proven useful in specific cases
  - Iterated grammatical learning winnows some patterns from the attested typology
  - Reanalysis shapes the relative frequency of attested patterns
- Modest support for winnowing, though not the whole story
  - Some unattested patterns remain; explanation for gap must lie elsewhere
- No clear support for learnability-based redistribution
  - In fact, r-volume remains marginally better predictor of typological frequency

What does a modeler do in this situation?

Wrong constraints?

- Hyde (2008, 2015): redefine ALIGN
- Kager (2012): reject *LAPSE because it can generate midpoint systems
  - Stanton (2016) addresses some but not all of these
- Switch to foot-based constraints?
  - Could eradicate ‘Penult $\leq$3σ else antepenult’ problem
  - May incorrectly rule out potential midpoint-type systems (Içũa Tupi, Bhojpuri)
  - Role of learnability far more difficult to explore: Hidden Structure (Tesar & Smolensky, 2000; Jarosz, 2013; Boersma & Pater, 2016)

R-volume as a prior?

- R-volume is not perfect, but striking that it does as well as it does. What creates this distribution?
- Bayesian inference: $P(H|D) \propto P(D|H) * P(H)$
  - Suppose ‘H’ is ‘grammar generating pattern *n*’ (aggregate over equivalent rankings)
  - Learners may retain H, even when likelihood (P(D|H)) is not especially high
  - Acoustic ambiguity of stress makes it especially prone to strong priors?
- Yang, Albright and Feldman (2022) Assessing the learnability of process interactions using grammatical spaces. (*Cogsci* paper)

A couple lessons

- The importance of scaling up
  - Prosthetic thought: calculating typology and r-volume, multiple runs of iterated learning
  - Results are not always what one expects
- Baselines
  - Results here aren't a particularly accurate model of typology, but they help point to some areas for future improvement
  - Empirical: accuracy of relative frequencies in StressTyp?
  - Attention to mismatches may point to changes in the constraints or learning model

Staubs (2014) on Stress windows

Stress windows

- Languages may have stress at a fixed distance from an edge, but not all distances are observed

Attested:	final	ό, σό, σσό, σσσό, σσσσό, σσσσσό
Attested:	penultimate	ό, όσ, σόσ, σσόσ, σσσόσ, σσσσόσ
Attested:	antepenultimate	ό, όσ, όσσ, σόσσ, σσόσσ, σσσόσσ
Unattested:	preantepenultimate (and beyond)	ό, όσ, όσσ, όσσσ, σόσσσ, σσόσσσ

Stress windows (*cont.*)

- Fixed Con solution: limit on constraints
  - *ExtendedLapse(R) penalizes όσσσ, prefers antepenultimate όσσ
  - Hypothesis: no equivalent *ExtendedExtendedLapse(R) (*όσσσσ)

/σσσσσσ/	*ExtLapse(R)	*ExtLapse(L)	Align(L)	Align(R)
a. όσσσσσσ	*! W		L	****** W
b. σόσσσσσ	*! W		* L	***** W
c. σσόσσσσ	*! W		** L	**** W
d. σσσόσσσ	*! W	*	*** L	*** W
☞ e. σσσσόσσ		*	****	**
f. σσσσσόσ		*	*****! W	* L
g. σσσσσσό		*	*****!* W	L

Staub's general claim

- Phonological grammar allows for window lengths of any(?) size
- However, longer window lengths are harder to learn, because the data needed to distinguish them from short windows is rare
    - Long windows show up only in long words (following Prince 1993, Pater 2009)
- Harder to learn = lower probability that an individual learner will acquire it successfully
- Iterated across generations: frequency of such patterns is reduced or eliminated

Illustrating the idea

- Staubs uses MaxEnt models with foot-based constraints, which intrinsically predict 4+σ windows
- To keep things consistent, I'll recast the problem into the constraints already presented above, adding an additional constraint to allow for 4σ windows
  - *3-Lapse(R/L): penalizes stressless σσσσ#, #σσσσ
  - With this notation, *ExtLapse = *2-Lapse

Amount of ranking data

/σσ/	*3-Lapse(L)	*3-Lapse(R)	*2-Lapse(L)	*2-Lapse(R)	Align-L	Align-R
a. όσ						1
b. σό					1	

/σσσ/	*3-Lapse(L)	*3-Lapse(R)	*2-Lapse(L)	*2-Lapse(R)	Align-L	Align-R
a. όσσ						2
b. σόσ					1	1
c. σσό					2	

/σσσσ/	*3-Lapse(L)	*3-Lapse(R)	*2-Lapse(L)	*2-Lapse(R)	Align-L	Align-R
a. όσσσ				1		3
b. σόσσ					1	2
c. σσόσ					2	1
d. σσσό			1		3	

/σσσσσ/	*3-Lapse(L)	*3-Lapse(R)	*2-Lapse(L)	*2-Lapse(R)	Align-L	Align-R
a. όσσσσ		1		1		4
b. σόσσσ				1	1	3
c. σσόσσ					2	2
d. σσσόσ			1		3	1
e. σσσσό	1		1		4	

Making the model sensitive to amount of data

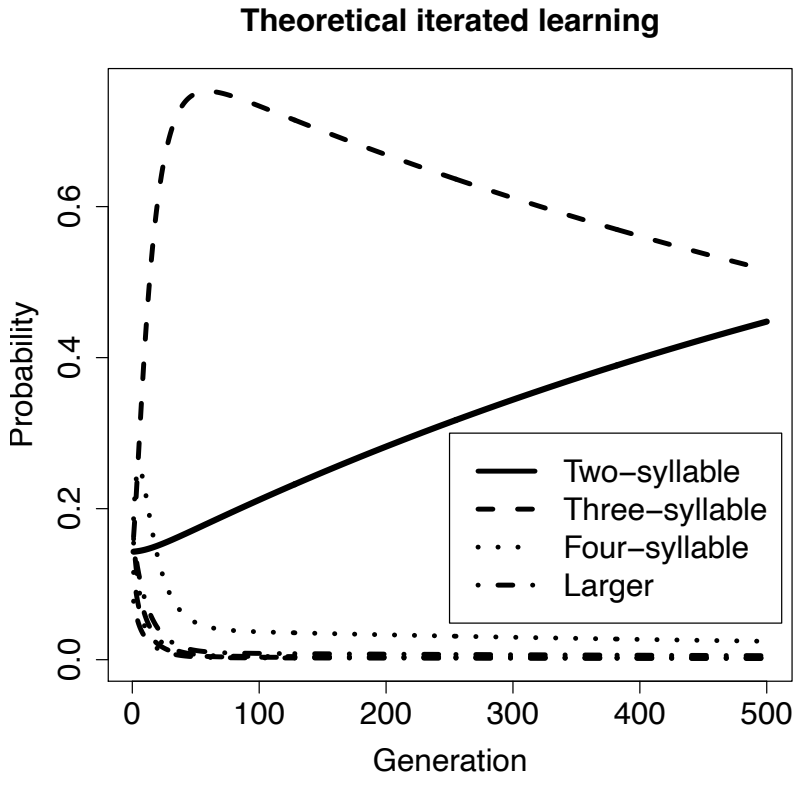
- Gradual learning: fewer relevant W/L pairs means fewer updates
- Priors/regularization: less data means the prior gets more of a say

Why would priors favor shorter windows?

- Given ambiguous data, why would learners favor shorter windows?
  - Train on 4σ window, but mislearn as 3σ window?
- 4σ window: *3-Lapse(R) ≫ Align-L ≫ *2-Lapse(R)
- 3σ window: *2-Lapse(R) ≫ Align-L
  - Ranking of *2-Lapse(R) doesn't matter
- Idea: if constraints are promoted, *2-Lapse(R) has more data favoring its promotion
- Even if data has words of all lengths, more inputs promote *2-Lapse(R)
  - 4σ and up for *2-Lapse, vs. 5σ and up for *3-Lapse

Staubs' result

Staubs (2014), Fig 2.8 (p. 96)



- As longer windows are reanalyzed, frequency of antepenultimate stress increases dramatically

References

Bane, Max, and Jason Riggle. 2008. Three correlates of the typological frequency of quantity-insensitive stress systems. *Proceedings of the tenth meeting of acl special interest group on computational morphology and phonology*, SigMorPhon '08, 29–38. Stroudsburg, PA, USA: Association for Computational Linguistics. Online: http://dl.acm.org/citation.cfm?id=1626324.1626330.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

References (*cont.*)

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, http://fon.hum.uva.nl/paul/.

Boersma, Paul, and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. *Harmonic serialism and harmonic grammar*, ed. by John McCarthy and Joe Pater, 389–434. Sheffield: Equinox.

Hyde, Brett. 2008. Alignment continued: distance-sensitivity, order-sensitivity, and the midpoint pathology. ROA 998.

Hyde, Brett. 2015. The midpoint pathology: What it is and what it isn't. ROA 1231.

References (*cont.*)

Jarosz, Gaja. 2013. Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology* 30.27–71.

Kager, René. 2012. Stress in windows: Language typology and factorial typology. *Lingua* 122.1454–1493.

Magri, Giorgio. 2012. Convergence of error-driven ranking algorithms. *Phonology* 29.213–269.

Riggle, Jason. 2010. Sampling rankings. University of Chicago ms.

Stanton, Juliet. 2016. Learnability shapes typology: the case of the midpoint pathology. *Language* 92.753–791.

References (*cont.*)

Staubs, Robert D. 2014. *Computational modeling of learning biases in stress typology*. University of Massachusetts, Amherst dissertation.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.