Computation, learning, and typology

Class 7: Substantive bias

Adam Albright and Eric Baković

CreteLing 2023 — July 2023

creteling2023.phonology.party

Taking stock

The recipe, so far...

Ingredients of a typology:

- Formal limits on patterns that can be stated
  - Expressive power of rules/constraints
  - Computational system (ordered rules, optimization of constraint violations, etc.)
  - Substantive limits: stipulations that further limit the typology
    - Limits on CON (exclude some constraints)
    - Metarankings (exclude some rankings)
  - Learning shapes grammar
    - Prefer grammars that are “simpler”, closer to “start state”, etc.
    - Availability of data: some grammars harder or impossible to motivate?

Goal today

- Mini demonstration of how these elements can come together to model typological distributions, with iterated learning
- Recasting biases and initial states as priors
    - Learning objectives that balance priors and fit
- Other ways of enforcing priors in constraint-based models

O’Hara (2021): stop place contrasts

The typology of stop place contrasts

- In principle, stops can contrast for a range of places—e.g., Toda (Dravidian)

p	labial
t̪	dental
t	alveolar
ʈ	retroflex

- Some languages have larger inventories of place contrasts than others
- Positional restrictions
  - Pre-vocalic (usually larger) vs. non-prevocalic (usually smaller)
  - Pre-consonantal (usually smallest) vs. absolute final position (can be somewhat larger)
  - Caveats abound—e.g., Toda bans t,ʈ in word-initial position (must be post-vocalic)

O’Hara (2021)

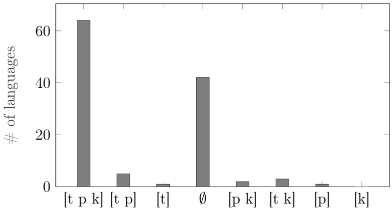
- 238 languages with (exactly) three-way p,t,k contrast
    - Collected from WALS, grammars
    - Represent 78 genera, 49 families
- For each, noted whether all three places available word-initially, word-finally

Onset/coda asymmetries

Among languages that allow p,t,k word-initially…

- Many also allow p,t,k word-finally
- Many allow no stops word-finally (“all-or-nothing”)
- Subsets: statistically, k $\Rightarrow$ p $\Rightarrow$ t

Figure 2.3: Word-final stop inventory given [t p k] initially



Onset/coda asymmetries

Among languages that allow p,t,k word-initially…

- Many also allow p,t,k word-finally
- Many allow no stops word-finally ("all-or-nothing")
- Subsets: statistically, k $\Rightarrow$ p $\Rightarrow$ t

Table 2.5: Languages in WALS with [t p k] initially.

Pattern	#	Language
All-Final	21	Abun, Alamblak, Asmat, Chamorro, Chontal Mayan, Cree (Plains), Daga, English, Georgian, Karok, Koasati, Korean, Kutenai, Lango, Ma'anyan, Meithei, Persian, Sierra Popoluca, Tagalog, Turkish, Yaqui
[tp]-Final	3	Indonesian, Kiowa, Oromo (Mecha)
[t]-Final	1	Finnish
No-Final	18	Apurinã, Arapesh (Mountain), Cubeo, Canela-Krahô, Fijian, Greek (Modern), Hixkaryana, Japanese, Kewa, Mandarin, Otomí (Mezquital), Pirahã, Quechua (Imbabura), Sanuma, Spanish, Supyire, Tukang Besi, Yagua
[pk]-Final	1	Lakhota
[tk]-Final	2	Imonda, Lavukaleve[b]

Capturing these asymmetries

- The initial/final asymmetry: positional faithfulness
  - IDENT(place)
  - IDENT(place)/ONSET

  (O'Hara's results don't actually speak to onset/coda; perhaps word-initial or pre-vocalic)
- Place asymmetries: a stringency hierarchy (de Lacy, 2004)
  - Markedness: *K, *KP, *KPT
  - Also: IDENT(K), IDENT(KP)

OT vs. MaxEnt

- O’Hara actually assumes weighted constraint grammars using MaxEnt, rather than strict rankings of OT
- Typologically significant properties of MaxEnt
  - Uses constraint weights to assign gradient probabilities to outputs
  - Constraints with equal weights can “tie”
  - Multiple violations of lower constraints can be worse than single violations of higher constraints
- Consequence: typology may be larger than factorial typology in OT
  - Infinite, once different probabilities of competing outputs are considered

The predicted typology

- 108 grammars yield 27 distinct patterns

Table 3.1: (Categorical) Patterns Predicted by Factorial Typology. Patterns highlighted in gray have less contrasting places of articulation than any attested language and may not have enough communicative power to be likely languages.

	Licit Forms								
	Name	tV	pV	kV	Vt	Vp	Vk	Attested?	Example
a.	No-Stops	✗	✗	✗	✗	✗	✗	✗	
b.	[t]-Initial	✓	✗	✗	✗	✗	✗	✗	
c.	[tp]-Initial	✓	✓	✗	✗	✗	✗	✓	Tahitian
d.	No-Final	✓	✓	✓	✗	✗	✗	✓	Barasano
e.	[t]-Final	✓	✓	✓	✓	✗	✗	✓	Finnish
f.	[tp]-Final	✓	✓	✓	✓	✓	✗	✓	Kiowa
g.	All-Final	✓	✓	✓	✓	✓	✓	✓	Tagalog
h.	Only-[t]	✓	✗	✗	✓	✗	✗	✗	
i.	[tp]-[t]	✓	✓	✗	✓	✗	✗	✗	
j.	No-Dorsals	✓	✓	✗	✓	✓	✗	✗	

Table 3.2: Gapped Patterns Predicted by Factorial Typology.

	Licit Forms								
	Name	tV	pV	kV	Vt	Vp	Vk	Attested?	Example
a.	[pk]-Final	✓	✓	✓	✗	✓	✓	✓	Korowai
b.	[tk]-Final	✓	✓	✓	✓	✗	✓	✓	Imonda
c.	[p]-Final	✓	✓	✓	✗	✓	✗	✓	Nimboran
d.	[tk]-Initial	✓	✗	✓	✗	✗	✗	✓	North Carolina Cherokee
e.	[pk]-Initial	✗	✓	✓	✗	✗	✗	✓	Hawaiian

R-volume

- Unlike OT, set of possible MaxEnt weightings is infinite
- Technique of Riggle (2010) can't be applied
- Instead, O'Hara (2021) uses sampling (1000 sets of weights) to estimate

W-volume estimate

Table 3.6: Predicted attestation rates based on R-volume

	Initial			Final			Percentage
No-Stops	✗	✗	✗	✗	✗	✗	22.9
[t]-Initial	tV	✗	✗	✗	✗	✗	26.6
[tp]-Initial	tV	pV	✗	✗	✗	✗	21.3
No-Final	tV	pV	kV	✗	✗	✗	14.1
[t]-Final	tV	pV	kV	Vt	✗	✗	3.7
[tp]-Final	tV	pV	kV	Vt	Vp	✗	1.8
All-Final	tV	pV	kV	Vt	Vp	Vk	.5
Only-[t]	tV	✗	✗	Vt	✗	✗	1.9
[tp]-Initial+[t]-Final	tV	pV	✗	Vt	✗	✗	3.1
No-Dorsals	tV	pV	✗	Vt	Vp	✗	.6

- All-final is most common typologically (>50%) but predicted to be very rare
  - Many constraints ban stops, few weightings put $\mathbb{F}$ above all of them
  - Unscrupulous fix: add many more copies of *Coda
- Some attested patterns predicted to be so rare that 1000 samples didn't find them

Not (only) simplicity

- The “all or nothing” split makes a ‘simple’ slice
  - Ban codas (or, at least, coda stops)
  - Could this reflect a preference to give higher weight to simpler/more general $\mathbb{M}$ constraints? (Albright & Hayes, 2006)
- O’Hara points out that there are other, equally simple slices that are rare or unattested
- Example: no dorsals
  - tV, pV, *kV, Vt, Vp, *Vk

Skews caused by learning?

- Recall Stanton (2016): some patterns are harder to learn than others, so predicted to be less stable
  - Further from start state
  - Data compelling reranking is rarer
- Could a similar effect be causing the observed typological skews?
- A careful test of this would start by analyzing discrepancies with typology, to see whether those languages are characterizable this way
  - O'Hara does do some of this

A quick attempt to test through simulation

- O’Hara compares what happens to different start states, sent through generational learning
- Goals
  - Are some patterns less stable than others?
  - Is probability redistributed to match the typology?
- Start state: flat distribution across patterns
  - Probably even worse assumption than r-volume distribution, but shouldn’t matter, given enough time?

Generational learning

- Generations: single learner, learns and then produces output to transmit to next generation
  - 3600 learning iterations, all inputs equally available (enough to learn well the first generation)
  - 25 generations
  - 100 runs per pattern, to sample what happens for each

Attempt 1: all constraints equal

A baseline (not shown by O’Hara)

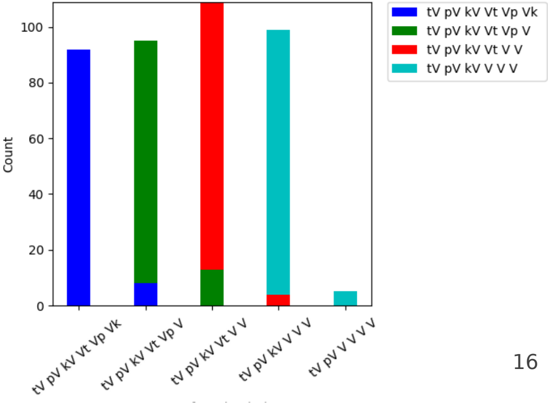
- All constraints start with initial weight of 50
  - Very large, but I’ll use it here because it’s what O’Hara uses
- Demo: SoftTypologyTool
- Results

Results: all constraints equal

- Recall relative proportions
  - All:21, [tp]-only:3, [t]-only:1, None:18
- Stability of different patterns

All-final	.9
[tp]-final	.8
[t]-final	.9
No-final	.9

- Redistribution



The model is 'too good' (and not good enough

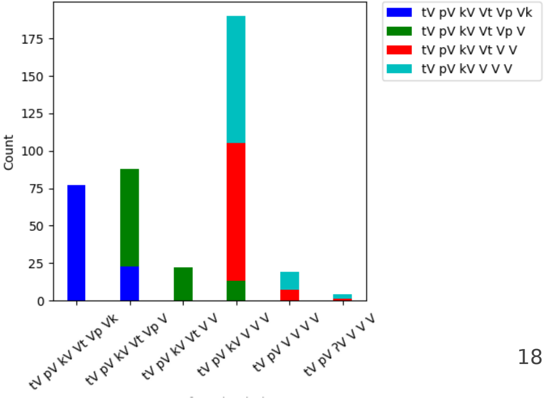
- Patterns generally learned quite well; infrequent patterns not unstable
- Faithfulness starts same as markedness, promoted quickly over all
- Resulting grammars generally allow even more structures than trained on (insufficiently restrictive)
  - Motivation for $\mathbb{M} \gg \mathbb{F}$ bias

Attempt 2: restrictiveness bias

- All $\mathbb{M}$ constraints start out the same (w=50), above faithfulness (w=1)
- Stability of different patterns

All-final	.7
[tp]-final	.6
[t]-final	.0
No-final	.8

- Redistribution



A model with no stringency hierarchy

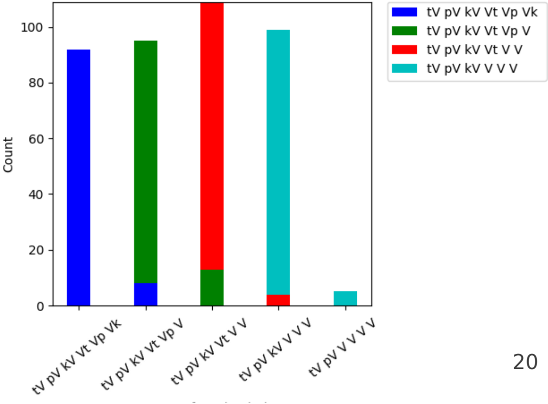
- O'Hara also compares a model that lacks the place markedness hierarchy
  - Can capture all combinations of places equally
  - *K, *P, *T, *KP, *KT, *PT, …
  - Also positive constraints +K, +P, +T, …
- Not really 'unbiased', but less asymmetry built into the model's biases

Results: no stringency hierarchy

- Stability of different patterns

All-final	.98
[tp]-final	.8
[t]-final	.9
No-final	.9

- Redistribution



Biases as priors

So far

- Bias encoded in constraint definitions
- Bias encoded in initial state of learning
  - “Default”, in the absence of data)
  - Brittle: immediately defeated by data
- In actuality, people balance default assumptions and evidence
- Biases are persistent

Biases as priors

- Priors $=$ beliefs about what is more or less probably, independent of the data
  - Before receiving any data
  - In spite of the data
- Not just initial states, but learning objectives

Incorporating priors

- Bayesian inference
  - P(H|d) = P(d|H)*P(H) / P(d)
- MaxEnt models
  - Rather than maximizing (log) likelihood of the data, maximize a complex term that combines data and priors

Priors for constraint-based grammars

Two commonly applied priors (first brought to phonology by Wilson, 2006)

- The weights of the constraints (μ)
- How much the weights of different constraints change in response to data (σ^2)

White (2017)

- Recall the Campdanian Sardinian problem from Class 6
  - p $\rightarrow$ β, but b $\neg\rightarrow$ β
- White's proposal
  - *Map(b$\sim$β) $\gg$ *VTV $\gg$ *Map(p$\sim$β)
- But such patterns are rare
  - Learning bias: *Map(p$\sim$β) $\gg$ *Map(p$\sim$b), *Map(b$\sim$β)

Bias against saltation

- Voiced stop lenition: p→p, b→β
  - Violates *Map(b∼β)
- Saltatory lenition: p→β, b→b
  - Violates *Map(p∼β)

Typologically common vs. rare patterns

Comparison set

p→p, b→b	No lenition
p→p, b→β	Voiced stop lenition
p→b, b→β	Chain shift
p→β, b→β	Total lenition
p→β, b→b	Saltatory lenition

A flat prior on *Map

- All *Map constraints start out weighted equally
- Any alternation is as likely as any other

A biased prior on *Map

- *Map(p∼β) higher than *Map(p∼b), *Map(b∼β)
- White (2013) shows that learners in the lab acquire saltation less accurately from data (AGL task)

Implementing this bias

White (2017): prior on weights (μ)

- Relative weights based on confusability in noise

*MAP(p∼v)	3.65
*MAP(f∼v)	2.56
*MAP(p∼b)	2.44
*MAP(f∼b)	1.96
*MAP(p∼f)	1.34
*MAP(b∼v)	1.30

- Markedness (*V[-voi]V, *V[-cont]V) = 0
- Consequence: higher prior on b→v than p→v

Typological implications

- $\mathbb{M} \gg \mathbb{F}$ prior should make faithful mappings less stable
- Lower prior on saltation mappings should make them less stable than non-saltatory mappings
- Various works on typology of lenition, but I'm not aware of any that counts all of these together(?)

Are these the right priors?

- Biases we contemplated, before talking about MaxEnt
  - Hard metarankings
  - Soft metarankings M»F, P-Map
- These are about relative rankings (*differences* in weights)

References

Albright, Adam, and Bruce Hayes. 2006. Modeling productivity with the Gradual Learning Algorithm: The problem of accidentally exceptionless generalizations. *Gradience in grammar: Generative perspectives*, ed. by Gisbert Fanselow, Caroline Féry, Ralf Vogel, and Matthias Schlesewsky, 185–204. Oxford University Press.

de Lacy, Paul. 2004. Markedness conflation in Optimality Theory. *Phonology* 21.145–199.

References (*cont.*)

O'Hara, Charles P. 2021. *Soft biases in phonology: Learnability meets grammar*. University of Southern California dissertation. Online: https://www.proquest.com/dissertations-theses/soft-biases-phonology-learnability-meets-grammar/docview/2535887726/se-2.

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.

White, James. 2013. *Bias in Phonological Learning: Evidence from Saltation*. UCLA PhD dissertation.

References (*cont.*)

White, James. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a p-map bias. *Language* 93.1–36.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science: A Multidisciplinary Journal* 30.945–982.