What Matters in Tonotactic Learning

Hán Li & Jeffrey Heinz

Stony Brook University
Department of Linguistics
Institute for Advanced Computational Science

SCiL 2026 July 3. 2026

Motivation

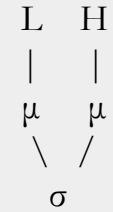
Why is tonotactic learning different?

Segmental Phonology

- Constraints are expressed **linearly**
  - Segments *CC
  - Features *[+nas][-voi, -cont]
- Current learning models:
  - **MaxEnt** (Hayes & Wilson, 2008)
  - **Neural models** (Mayer & Nelson, 2020)
  - **Probabilistic Models** (Coleman & Pierrehumbert, 1997; Goldsmith & Riggle, 2012, a.o.)
  - **BUFIA** (Chandlee et al., 2019)
- Linear representations fit naturally into existing learning algorithms

Tonal Phonology

- Constraints are expressed **autosegmentally**
  - Multi-tier structures
  - Association relations



- Current learning models:
  - **flatten** tones into strings (Wiebe, 1992; Bird & Ellison, 1994; Shih & Inkelas, 2013; Rawski & Dolatian, 2020)
  - Directly operates AR (**BUFIA-AR**, Li, 2025; Li & Heinz, to appear)
- **Do ARs improve tonotactic learning?**

In This Talk

We ask two questions:

- Does **representation** affect learning performance?
- Does the answer depend on the learning **schemes (model, frequency, complexity)?**

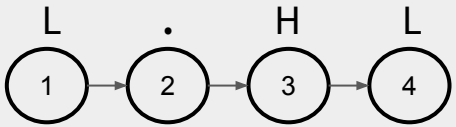
To answer these questions, we compare three tonal representations and two learning algorithms (MaxEnt & BUFIA) using a Hausa corpus. Our results show:

- **BUFIA-AR** achieves ceiling performance with sparse data and no freq.
- No consistent advantage between segmental and feature representation.
- Frequency information substantially improves MaxEnt performance on seg/ftr.
- Increasing model complexity leads to overfitting.

Three Model-Theoretic Representations

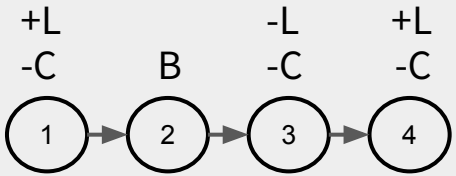
***mù.tûm* "person" (Hausa)**

Segment



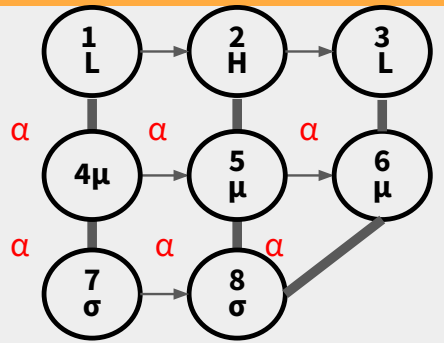
L, H	Low/High
.	syllable boundary
→	Linear order

Feature



+L, -L	Low/High
+C, -C	contour/level
B	syl boundary
→	Linear order

ARs



L, H , μ, σ	L, H, mora, syl
→	Linear order
α	**Association**

Two Learning Models

MaxEnt (Hayes & Wilson, 2008)

- Maximum Entropy Model
- A Probabilistic learner
- Grammar of weighted constraints
- Supports segment and feature

BUFIA (Chandlee et. al., 2019)

- Bottom Up Factor Inference Algorithm
- A Categorical learner
- Grammar of Forbidden factors
- Support model-theoretic representation

	Segment	Feature	AR
MaxEnt	✅	✅	❓
BUFIA	✅	✅	✅

Experiment

Does Representation Matter?

Language: Hausa

- Hausa: West Africa
- 3 tones: H, L, F (HL)
- General constraints:

Constraints	Interpretation
(1) *RISE$_\sigma$	No monosyllabic LH
(2) *CONTOUR$_\mu$	No contour on single mora
(3) *3T	No trimoraic contour
(4) *NONFINHL	No non-final HL
(5) *LAPSE	No adjacent L-toned syllables

- Corpus: Hausa WOLD (Awagana, 2009)
  - **621** words
  - **64** distinct patterns

4-fold cross validation

Training

- Positive data only (¾ of lexical **types**)

Word	String	Feature		AR
mù.tûm 'person'	L.HL	L:	+0−+	L H L | | | μ μ μ | \ / σ σ
		C:	−0−−	
		B:	0+0 0	

- Train on models
  - MaxEnt (k=3)
  - BUFIA-AR (t = 2, s = 2, m = 3)

Testing

- Held-out set (¼ types)
- Negative data:
  - Self-generated (same size)

(nonce)	R.H	L:	+0−	L H | ⁄ | μ μ | | σ σ
		C:	+0−	
		B:	0+0	

- Evaluate on
  - **Baseline**
  - MaxEnt
  - BUFIA
- Prec, Recall, F1

Baseline Grammar across Reps

(Newman, 2002; Leben, 1996; Zoll, 2003).

Constraint	String	Feature			AR
*RISE$_\sigma$	LH	L:	+	−	
		C:	−	−	
		B:	0	0	
*CONTOUR$_\mu$	R	L:	+		
		C:	+		
		B:	0		
	F	L:	−		
		C:	+		
		B:	0		
*NOFINHL	HL.	L:	−	+ 0	
		C:	−	− 0	
		B:	0	0 +	
*LAPSE	L.L	L:	+	0 +	
		C:	−	0 −	
		B:	0	+ 0	

*3T	HHH	L:	−	− −	
		C:	−	− −	
		B:	0	0 0	
	LLL	L:	+	+ +	
		C:	−	− −	
		B:	0	0 0	
	HHL	L:	−	− +	
		C:	−	− −	
		B:	0	0 0	
	HLL	L:	−	+ +	
		C:	−	− −	
		B:	0	0 0	
	HLH	L:	−	+ −	
		C:	−	− −	
		B:	0	0 0	

Results

- **Baseline**
  - Identical performance across all three reps
  - Rep alone does not affect a fixed grammar.
- **BUFIA**
  - Performance depends on representation.
  - Segment = Feature (F1 = 0.992, near ceiling)
  - **AR performs best (F1 = 1.000).**
- **MaxEnt:**
  - Representation also matters.
  - Segment (F1 = 0.871) > Feature (F1 = 0.832)
- **BUFIA achieves higher** performance than MaxEnt, although both models are near ceiling on this dataset.

Model	Recall	Precision	F1
Base-Seg	0.812	1.000	0.894
Base-Ftr	0.812	1.000	0.894
Base-AR	0.812	1.000	0.894
BUFIA-Seg	0.984	1.000	0.992
BUFIA-Ftr	0.984	1.000	0.992
BUFIA-AR	1.000	1.000	1.000
MaxEnt-Seg	0.922	0.827	0.871
MaxEnt-Ftr	0.906	0.773	0.832

Table 4: Average model performance

Follow-up Question 1

Does frequency matter?

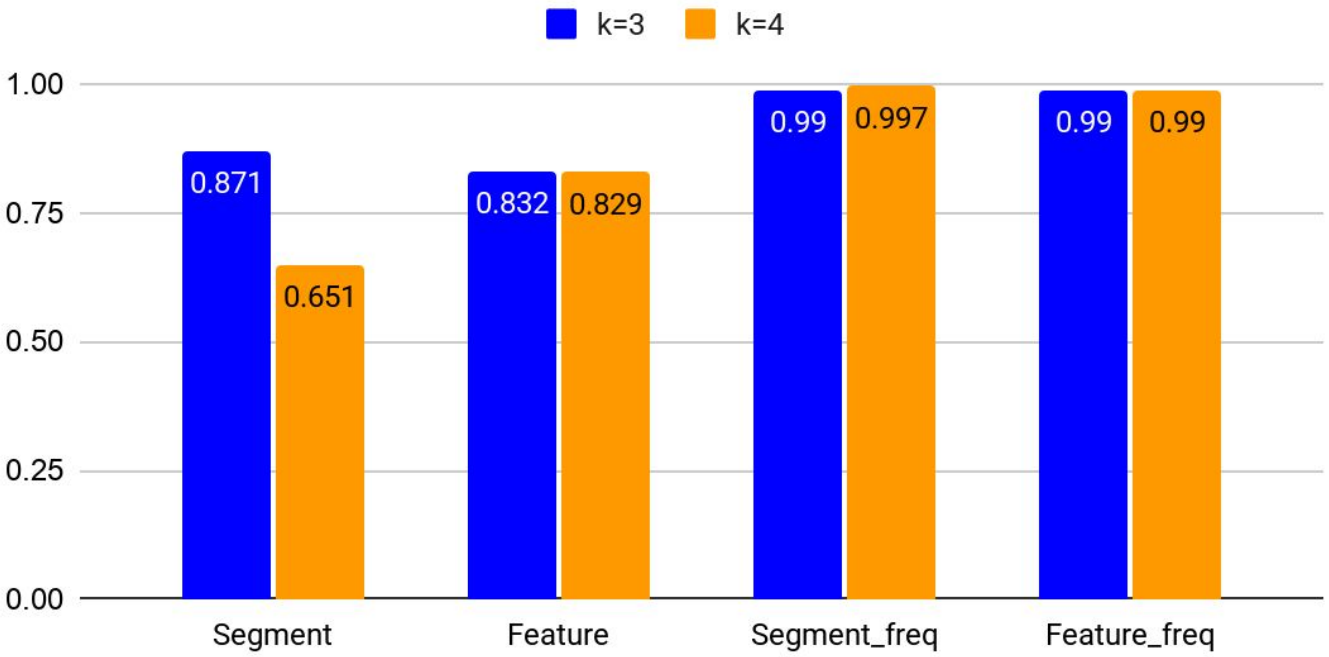
- **BUFIA**
  - A categorical learner does not depend on frequency information.
  - Performance stays the same across representations
- **MaxEnt:**
  - A Probabilistic learner benefits substantially from token frequencies
  - Seg: F1 = 0.871 → 0.990
  - Ftr: F1 = 0.832 → 0.990
- **Frequency, not representation, is the main limiting factor for MaxEnt.**
- **With frequency information, MaxEnt-Seg and MaxEnt-Ftr perform equally well, near ceiling.**

Follow-up Question 2

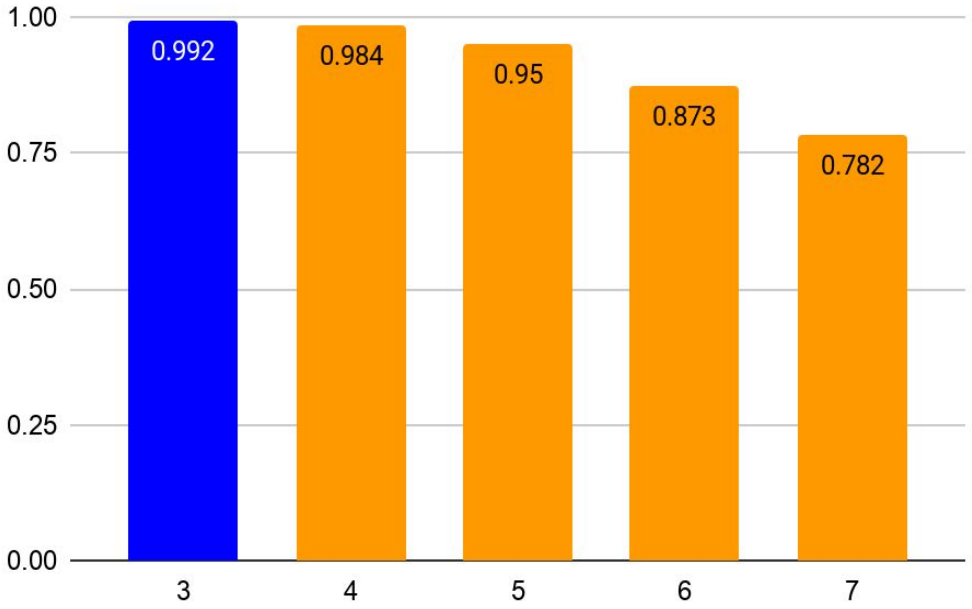
Does complexity matter?

- **Increasing model complexity leads to overfitting.**
- **MaxEnt: (F1, without freq)**
  - Seg: 0.871 (k=3) → 0.651 (k=4)
  - Ftr: 0.832 (k=3) → 0.829 (k=4)
- **BUFIA**
  - Seg/Ftr: F1 = **0. 992** → **0. 984** → **0.950** → **0.873**
  - AR: F1 = 1 (t = 2, s = 2, m = 3) → **0. 992** (t = 2, s = 2, m = 3) → **0. 967** (t = s =m = 3)

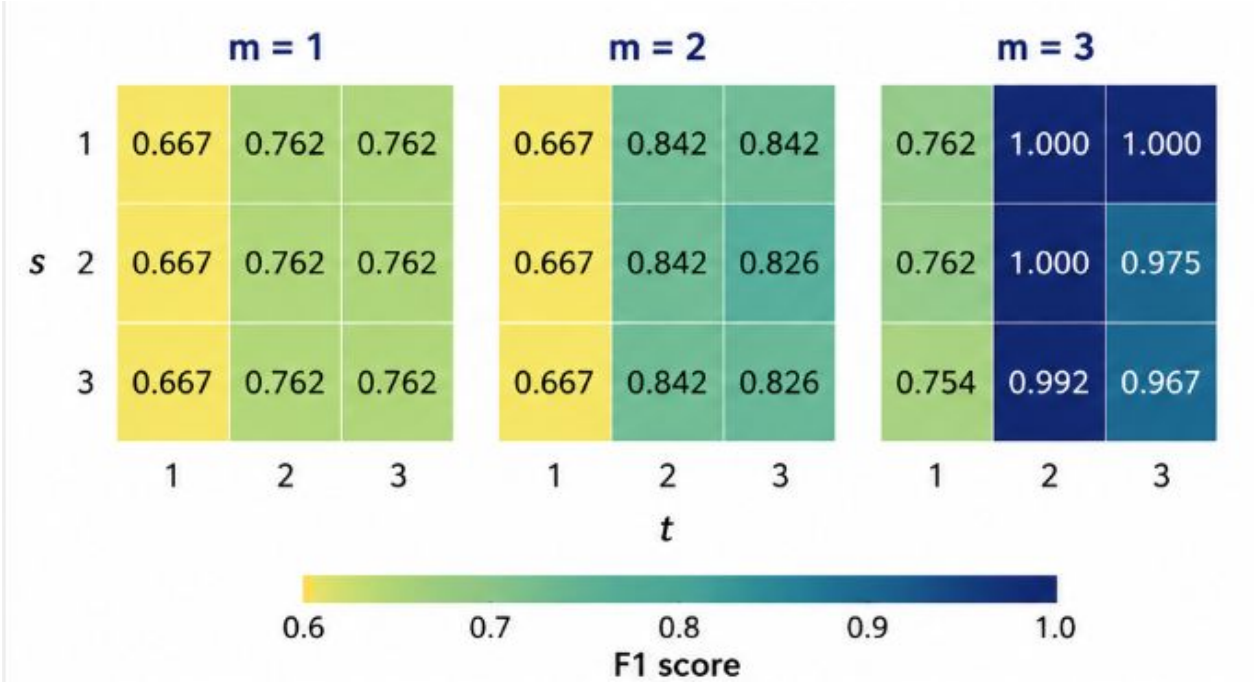
MaxEnt F1 with freq * k



BUFIA-Seg/Ftr F1 with varying k



BUFIA-AR F1 with varying (t,m,s)



Conclusion

What have we learned for tonotactic learning?

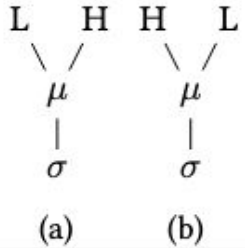
- **BUFIA-AR achieves the best** performance with sparse data and no frequency information, though with other representations are near ceiling in some settings.
- MaxEnt requires **frequency** information for successful tonotactic learning.
- Increasing model complexity leads to overfitting.

Limitation and Future work

- Experiments were conducted on a single language (Hausa)
- Only two types of learning models were evaluated.
- Future work should evaluate additional learning models (Neutral Networks) and more tonal systems.

Backup slides

BUFIA-AR Learned Grammar

	$\mu = 1$	$\mu = 2$	$\mu = 3$
$s = 1$	 (a) (b)	 (c)	 (d)
$s = 2$	-	-	 (e)  (f)

MaxEnt-Ftr Learned Grammar

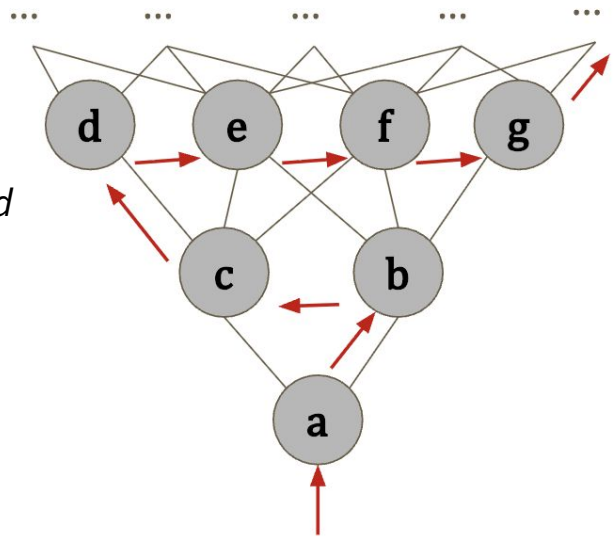
- *[+R]
- *[+F]
- *[+B][+B]
- *[+L][+H]
- *[+word_boundary][+B]
- *[+B][+word_boundary]
- *[+H][+L]
- *[+H][+H][+H]
- *[+L][-word_boundary][+L]
- *[-word_boundary][+H][+L]
- *[+B][-word_boundary][+word_boundary]
- *[+L][+L][+L]
- *[+H][+L][+L]
- *[+H][+L][-word_boundary]
- *[+word_boundary][+L]
- *[+B][+H]
- *[+word_boundary][-word_boundary][+B]
- *[+B][-word_boundary][+B]
- *[+H][+word_boundary]
- *[+H][-word_boundary][+H]

BUFIA

Input: A set of positive data D (attested forms)

Search: Start with the the most general structure s:

- If s is *eligible* (under a certain complexity level) and *attested* (contained in D)
  - generate its next superfactors
  - check the superfactors in turns
- If s is unattested:
  - add s to Grammar
  - do not expand it further (pruning off its superfactors)



Output: A set of inviolable constraints that are consistent with all attested data.

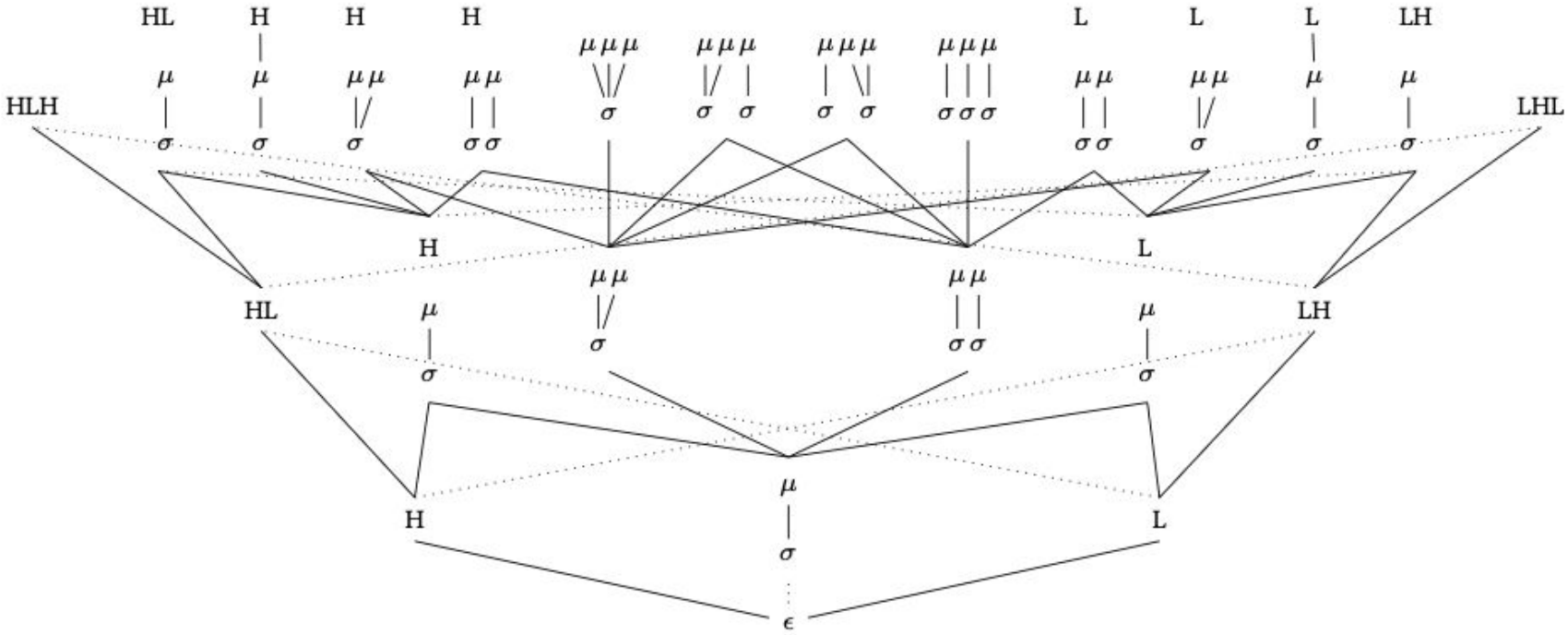
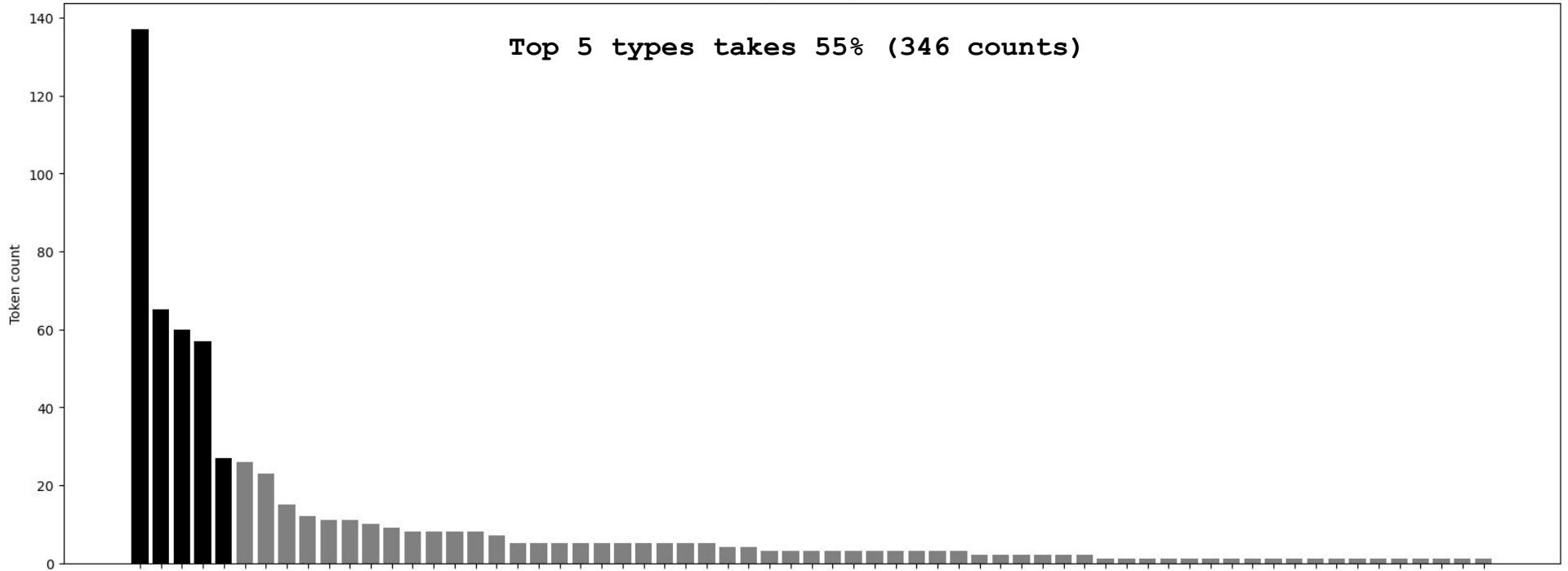
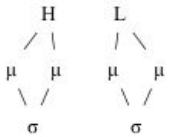


Figure 9: A partially-ordered hypothesis space over ARs

Data Distribution



The Five Most Common ARs



HH.LL

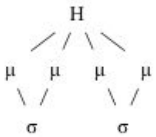
138

22.06%

H.LL

65

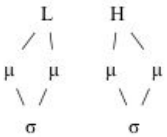
10.47%



HH.HH

60

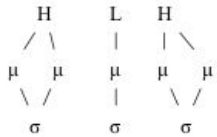
9.66%



LL.HH

57

9.18%



HH.L.HH

27

4.35%