

1. Overview

Language specific factors (phonetics, lexical frequency, etc.) can be associated with particular grammatical patterns.

- **Contour Tone Licensing** patterns with language-specific duration properties (Zhang 2002), and syllable type frequency (here).

Two ways to model a typological association in constraint-based frameworks

- **Direct:** The language-specific factor is encoded directly into the constraint set.

Possible with duration, but challenging with frequency.

- **Learning:** The language-specific factor influences the learnability of patterns, indirectly leading to the association.

Frequency affects learnability (and phonetics might too).

CLAIM: Learning can capture associations between language-specific factors and grammatical patterns.

- Syllable type frequency is associated with contour tone licensing patterns.
- Frequency conditions learnability, matching observed association.

2. Contour Tone Licensing

Multiple disjoint factors can influence contour tone licensing.

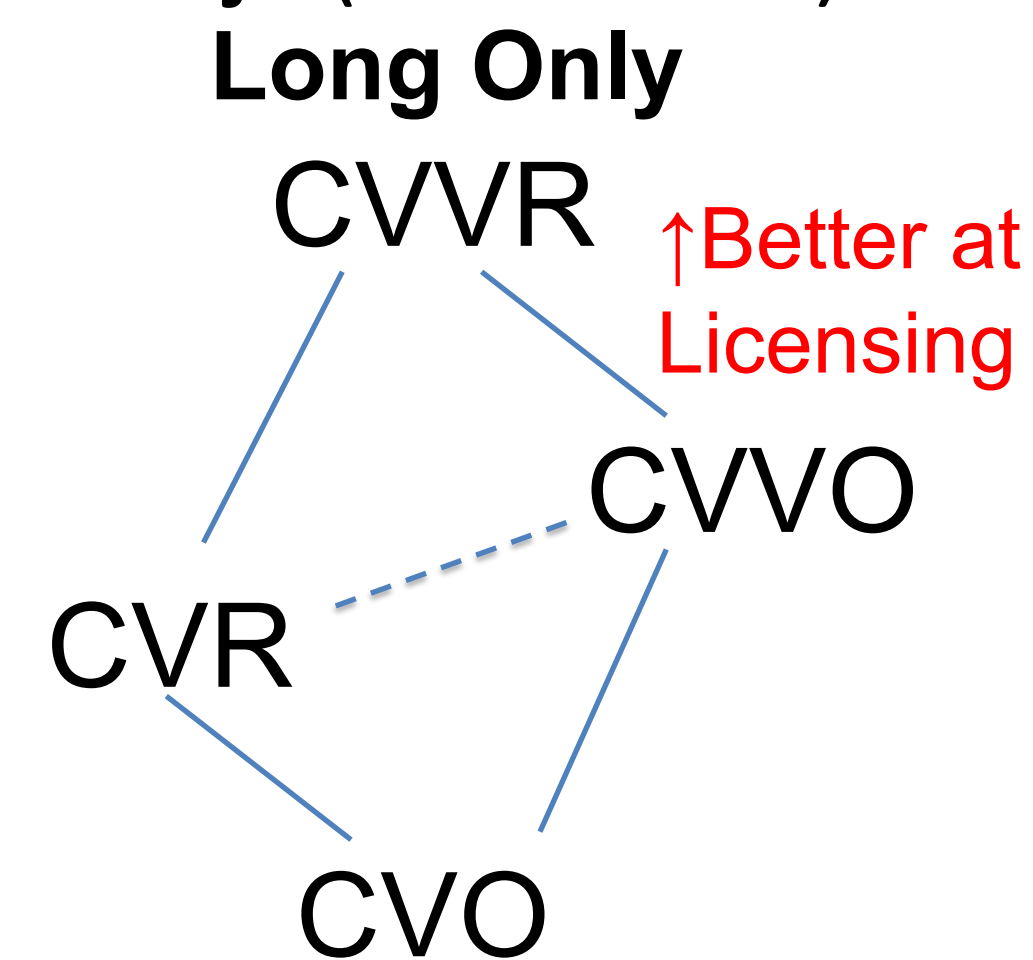
Vowel Length (Zhang 2002, 2004)

Long > Short
CVV(C) > CV(C)

Coda Sonority

Unchecked > Checked
CV(V)(R) > CV(V)O

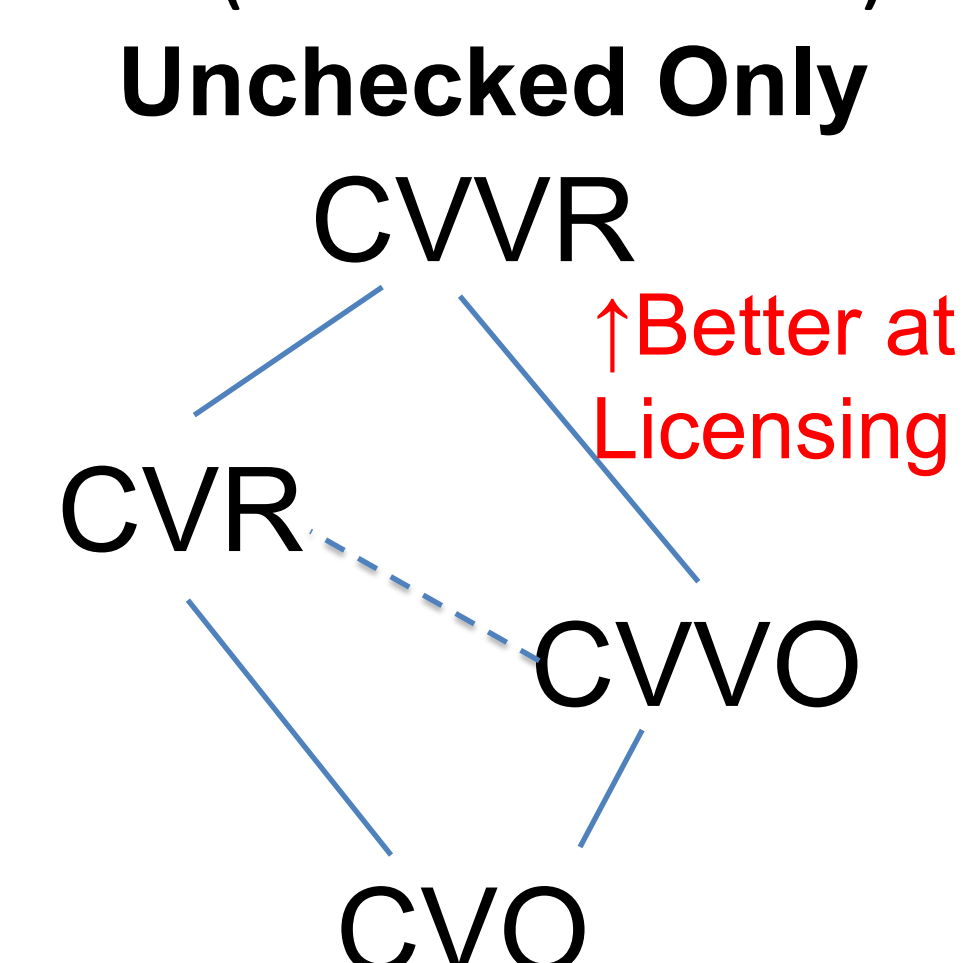
In **Navajo** (and Somali) contours on **long vowels only**.



Navajo	CV(V)	CV(V)O	CV(V)R
Short	*[sáni]	*[pitít]	*[pik'in]
Long	[há:kò:nè:ʔ] 'let's go'	[tè:ɬɬɬnì:ton] 'they shot at him'	[tè:ɬɬɬá] 'they extend'

(Zhang 2002)

In **Thai** (and Cantonese) contours on **unchecked syllables only**.



Thai	CV(V)	CV(V)O	CV(V)R
Short	N/A	*[lǎk]	[lǎŋ] 'back'
Long	[nǎ:] 'thick'	*[lǎ:k] 'grandchild'	[lǎ:ŋ]

(Morén & Zsiga 2006)

3. Corpus Study

I extracted the type frequency of each syllable type from corpora of Navajo and Thai.

Navajo:

Frequencies from 39,767 lemmas extracted from Wiktionary (Cotterell et al 2017).

- Only final syllables were counted

Thai

Frequencies from 2,961 words of child-directed speech I extracted from CRSLP-MARCS on Childe (Luksaneeyanawin 2000).

Navajo	CV(V)	CV(V)O	CV(V)R
Short	25%	25%	1%
Long	11%	37%	1%

Thai	CV(V)	CV(V)O	CV(V)R
Short	N/A	12%	25%
Long	27%	13%	23%

Navajo

Checked Syllables = 62%

Short Syllables = 51%

Thai

Checked Syllables = 25%

Short Syllables = 37%

Navajo has more **checked syllables** than **short syllables**.

Thai has more **short syllables** than **checked syllables**.

4. Learning Model

Generational MaxEnt-learner to uncover learning bias (Staub 2014, Dowman et al 2006, Hughto 2018)

- Learners initialized: markedness high, faithfulness low.
- Limited amount of training data per generation.
- Resulting grammar trains next generation.
- Harder to learn → more likely to change across generations.
- Stability across 40 generations over 50 runs as learning metric.

On each iteration, the learner was exposed to a form from the target grammar, sampled according to the lexical frequencies.

• **Three Lexical Frequency Conditions**

- *Navajo, Thai and Control*

• **Two Target Grammar Conditions**

- *Long Only, and Unchecked Only*

Constraints Used:
Ident-Tone, *Contour,
*Contour/Short,
*Contour/Checked

Long only	CV(V)	CV(V)O	CV(V)R
Short	[CV]	[CVO]	[CVR]
Long	[Cǎ:]	[Cǎ:O]	[Cǎ:R]

or

Unchecked only	CV(V)	CV(V)O	CV(V)R
Short	[Cǎ]	[CVO]	[CǎR]
Long	[Cǎ:]	[Cǎ:O]	[Cǎ:R]

Sample according to Frequency Condition



Stability: Output grammar at Learner 40 = Original Target Grammar

5. Simulation Results

- Lexical frequency conditions which pattern is more learnable.

Stability	Long Only	Unchecked Only	Average Iterations	Long Only	Unchecked Only
Control	22%	24%	Control	2200	2200
Navajo	94%	10%	Navajo	1800	2400
Thai	0%	70%	Thai	>3000	1900

- Stability is correlated to the number of iterations needed on average to reach 95% accuracy.

- **Long Only:** checked syllables > short syllables
- **Unchecked Only:** short syllables > checked syllables

- Over generations, languages where the lexical frequency mismatches the pattern are more likely to be unstable.

6. Conclusion

Correlation between lexical frequencies and contour licensors emerges from grammar w/o constraints referring to frequency.

Language specific properties can influence typology through learnability rather than being directly encoded in the grammar.

- Looked at Frequency here, but other language-specific factors would influence learning as well.

FUTURE WORK: Phonetics skew observed input frequencies: Low duration syllables with contour tones are more likely to be misperceived/reduced.

- This channel bias affects the data presented to the learner, skewing the distribution towards longer duration forms.

