How Statistical Learning Impacts the Sound Patterns of the World's Languages

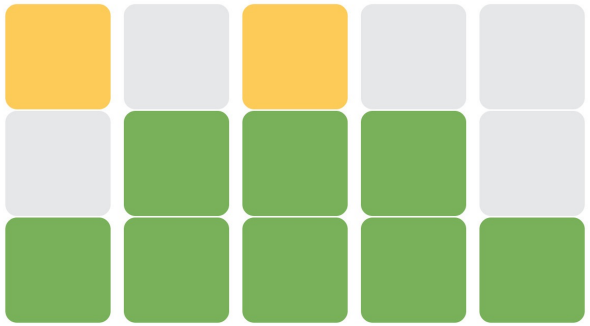
Charlie O'Hara

University of Michigan

1/19/22

Introduction

- Speakers have unconscious knowledge of properties of languages they speak
  - Not all sounds can appear in any position in a particular language
  - No “ng” /ŋ/ at the beginning of words in English
- **Phonotactics** – the language-specific rules that govern which sounds can appear in which contexts.



Word-Final Consonants

- Phonotactic knowledge affects how foreign words are borrowed into languages.

English

Internet [ɪntɹnɛt]
Jeep [dʒip]
fake [feɪk]

Polish

spłat	[spwat]	‘payoff’ gen pl
łap	[wap]	‘paw’ gen pl
rak	[rak]	‘cancer’

Finnish

olut	[ˈo.luṯ]	beer
**olup*	*[ˈo.lup]	✗
**oluk*	*[ˈo.luk]	✗

Japanese

インターネット	*intānetto*	[ĩntą̄ːnẹ̄t̪̄t̪o̞]	✗
ジープ	*jīpu*	[d͡ʑiːpɯ̟ᵝ]	✗
フェイク	*feiku*	[ɸẹ̟ːkɯ̟ᵝ]	✗

Word-Final Consonants

- Phonotactic knowledge affects how foreign words are borrowed into languages.

English

Internet [ɪntɹnɛt]
Jeep [dʒip]
fake [feɪk]

Polish

internet [ˈin.ter ˌnɛt] ✅
dżip [d͡ʒip] ✅
fejk [fɛjk] ✅

Finnish

internet [ˈintˌerneˌt] ✅
jeeppi [ˈjeːpːi] ❌
feikki [ˈfeikːi] ❌

Japanese

インターネット *intānetto* [ĩntaːneˀtˀto] ❌
ジープ *jīpu* [d͡ʑiːpɯᵝ] ❌
フェイク *feiku* [ɸeːkɯᵝ] ❌

Word-Final Consonants

- Some of these patterns are common, but some are very rare

English

Internet [ɪntɹnɛt]
Jeep [dʒip]
fake [feɪk]

Polish (t p and k)

Abun, Aklan, Alamblak, Apinaje, Arara, Asmat, Barok, Cebuano, Cree, Daga, Georgian, Korean, Lango, Persian, Tagalog, Turkish, Yaqui

Finnish (only t)

No other languages

Japanese (not t p or k)

Adamawa Fulani, Apalai, Apurinã, Arapesh, Canela-Krahô, Fijian, Greek, Hixkaryana, Kalapalo, Mandarin, Otomì, Pirahã, Quechua, Spanish, Tibetan, Warekana

Introduction

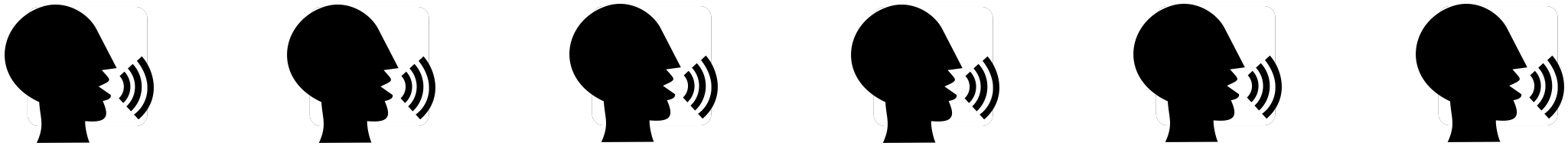
- Why are some patterns more common than others?

My work argues that **learnability** impacts the frequency of linguistic patterns.

Common patterns are **easy-to-learn.**

Learning (for this talk)

- Using computational models of phonological acquisition
- Idealizing the learning environment
  - A child receives input from one parent
  - A parent speaks one "language" using one grammar
  - After "hearing" a lot of words, the child stops learning, and becomes the parent of a new child.



What makes something hard-to learn?

Here, I focus on two case studies, though there are many other factors that affect learnability.

- Patterns that can be defined more **generally** are easier-to-learn than those with specific restrictions.
  - Word-Final Consonant Inventories
- Patterns that have more restrictions on structures that are **rare** in the lexicon of the language are easier-to-learn than those that restrict common structures.
  - Contour Tone Licensing

General patterns are easier-to-learn

(O’Hara 2021, in review)

Word-Final Stops (t, p, and k)

- In English, all three are allowed both at the beginning and the end of words.

[ti]	[pɑ]	[kɔ]
tea	*pa*	caw
[it]	[ɑp]	[ɔk]
eat	opp	awk

Word-Final Stops (t, p, and k)

- In Italian, all three are allowed at the beginning but **not** at the end of words.

[**t**asto]	[**p**asto]	[**k**asto]
button	*meal*	chaste
*[kasa**t**]	*[kasɑ**p**]	*[kasa**k**]

Word-Final Stops (t, p, and k)

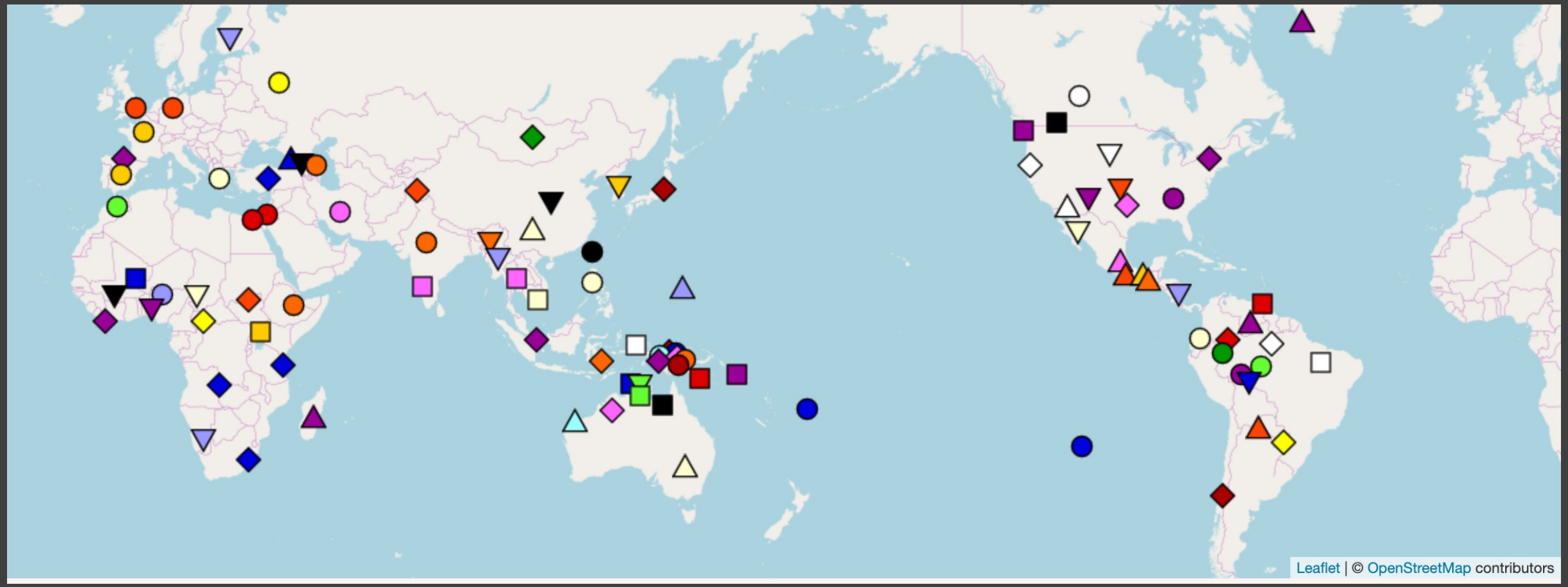
- Some languages allow only a subset of the stops word-finally
- In Movima (Bolivia), only t and p are allowed at the end of words.

[**t**anna]	[**p**ɛnna]	[**k**anan]
I cut	*my landing place*	your food
[tʃu:ha**t**]	[ku:du**p**]	*[ku:du**k**]
palm tree	*flea*	

Word-Final Stops (t, p, and k)

- Some languages allow only a subset of the stops word-finally
- In Finnish, only t is allowed at the end of words.

[**t**elata]	[**p**elata]	[**k**elata]
to paint with a roller	*to play*	to wind
[keo**t**]	*[keo**p**]	*[keo**k**]
anthills		



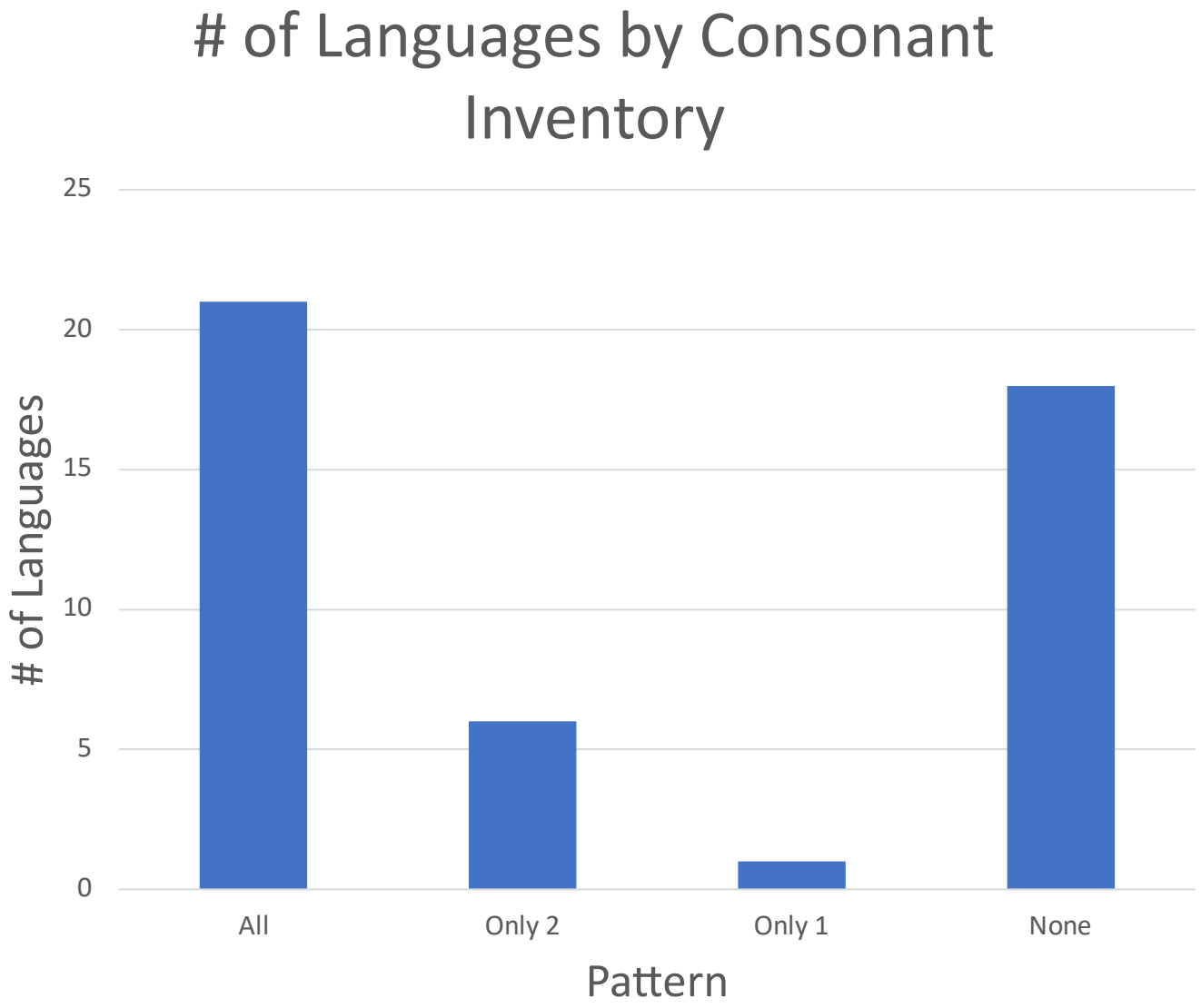
Which Patterns are Common?

(O'Hara 2021, in review)

- I investigated whether [t p k] appeared at the end of words in a sample of 94 languages (Dryer and Hapselmath 2014).
- I focus on a subset of 45 languages that avoid confounding factors.

Soft Typology of [t p k] at the end of words

- Languages tend to allow either all three, or none of [t p k] word-finally (88%)
- Subsets of [t p k] are rare.



Learning Simulation: Learning Agents

- Each learning agent has a MaxEnt Harmonic Grammar.
- Maxent assigns a probability to input-output mappings (x→y) based on a set of positive weights on a set of **constraints** (*features* in Comp. Sci.)
- The more higher weighted constraints a mapping *violates,* the less probable the mapping is.

to pe ka

ot ep ak

	8	4	3		
/ak/	Don't Delete	No Final Consonants	No [k]	Harmony Score	Probability
[ak]		-1	-1	-7	.73
[a]	-1			-8	.27

$$H(x,y) = \sum_{c \in \text{CON}} w_c * C(x,y)$$

$$P(y|x) = \frac{e^{H(x,y)}}{\sum_{z \in \text{Cand}(x)} e^{H(x,z)}}$$

Learning Simulation: Learning Algorithm

- Learners learn via a Stochastic Gradient Ascent algorithm.
- "Parent" and "child" both choose output forms *y* for a random input *x*

x

y_p Parent

y_c Child

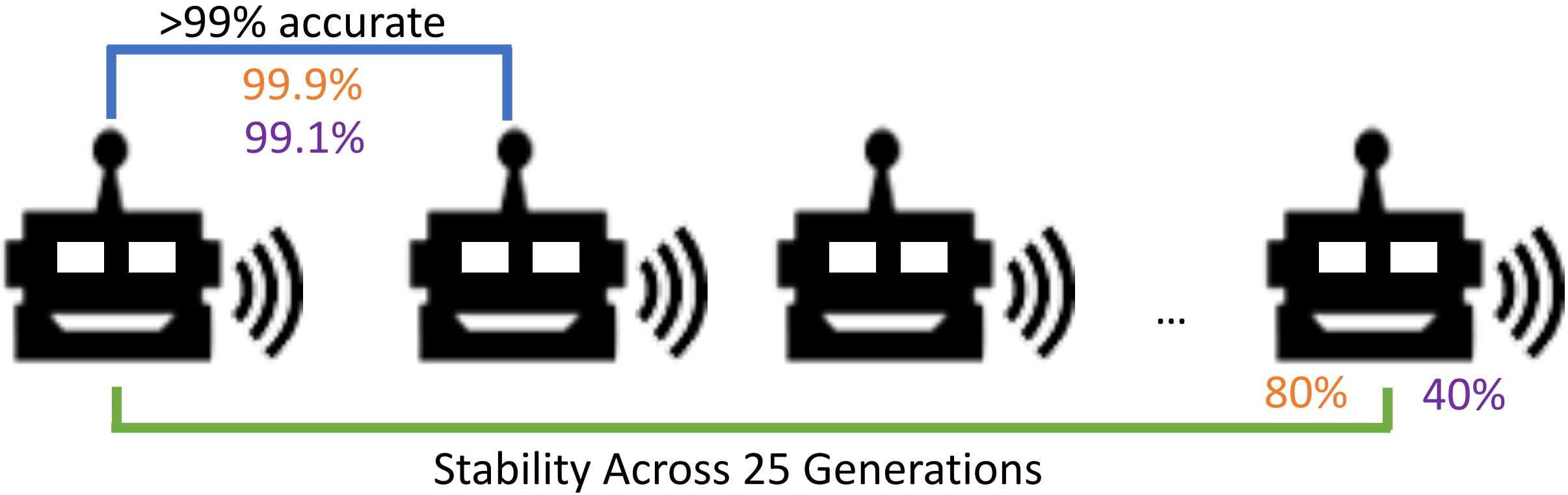
[ak]	/ak/	[a]
[ka]	/ka/	[a]
[ep]	/ep/	[e]
[pe]	/pe/	[e]

$$Update\ Rule$$
$$\Delta w_C = \mu(C(x, y_p) - C(x, y_c))$$

	10 ← 9	3	3		
/ak/	**Don't Delete**	No Final Consonants	No [k]	Harmony Score	Probability
[ak]		-1	-1	-6	0.982
[a]	-1			-10	0.018

Learning Simulation: Generational Learning

- This algorithm *weakly converges,* but human lives are finite
- Large, but limited number of forms per generation
- Easier to learn patterns are more stable than harder to learn patterns.

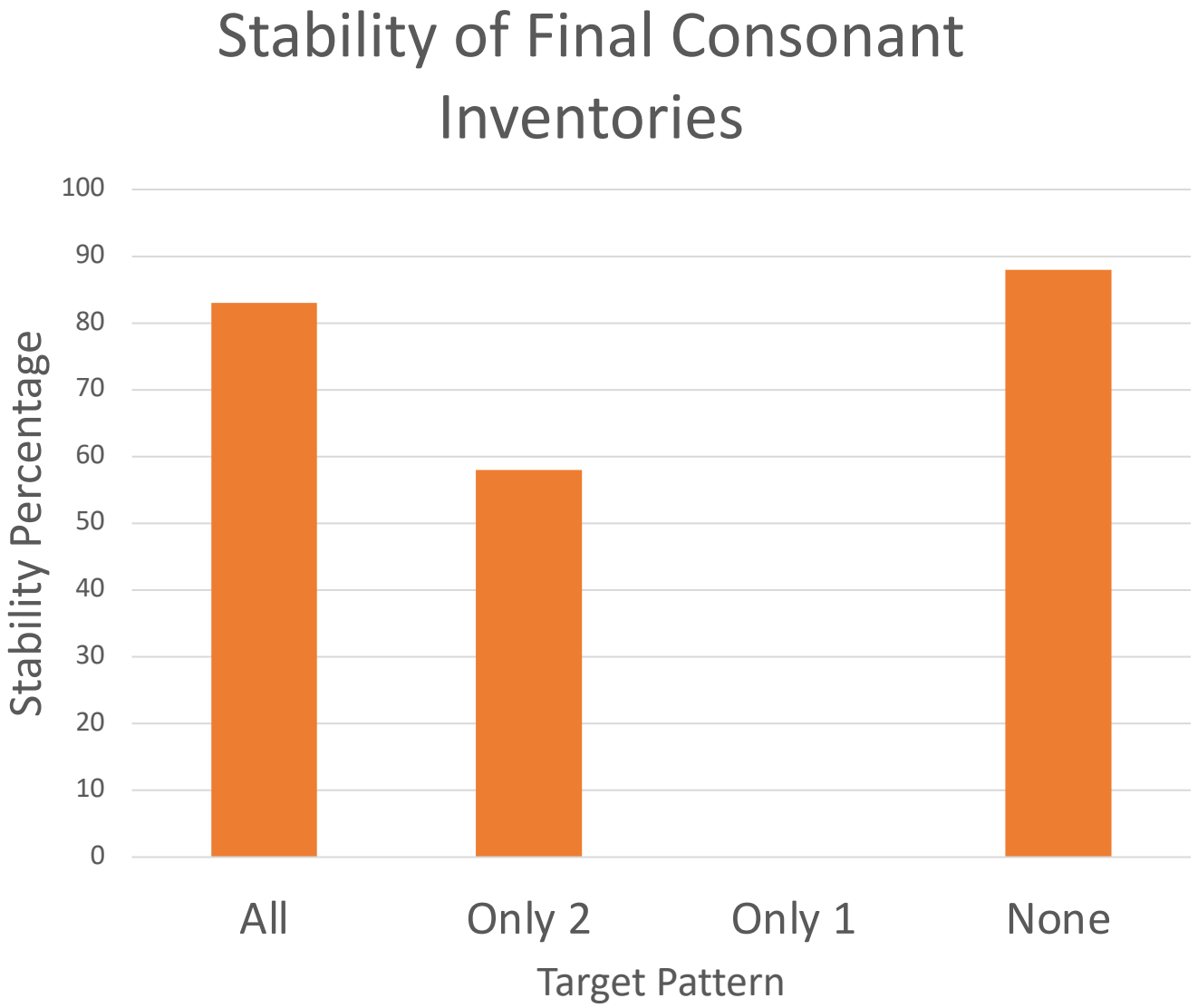


How learnable are different phonotactic patterns?

100 runs were done for each of the four patterns.

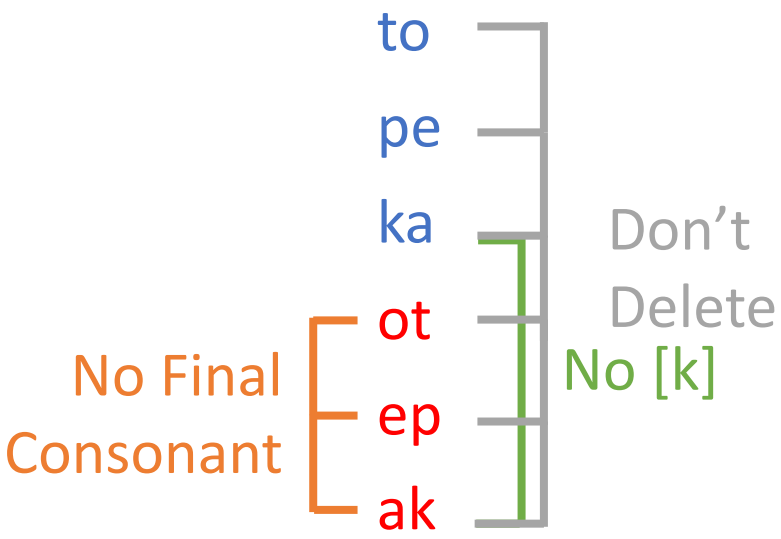
The most stable patterns are those that allow all or none of [t p k]

Subset patterns are less stable across generations.



Why are some patterns more learnable?

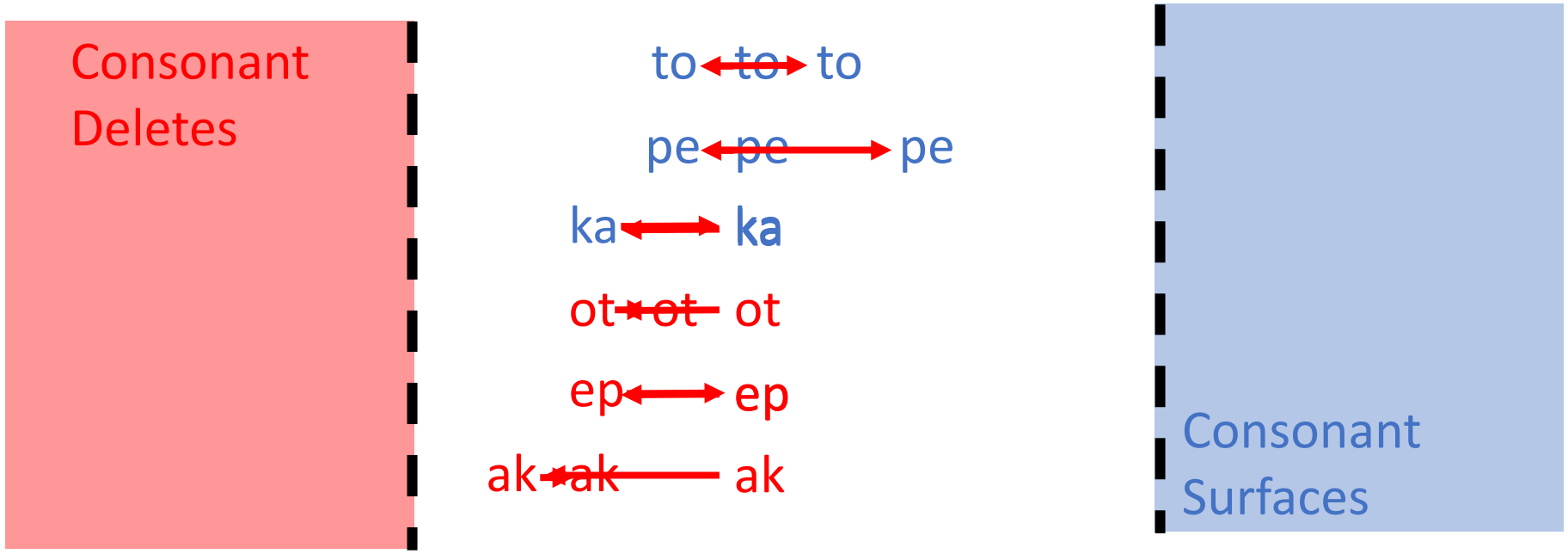
- The learnability of patterns is based on the constraints used to distinguish forms.
- Patterns that use general constraints consistently are easier to learn than those that do not.



Why are some patterns more learnable?

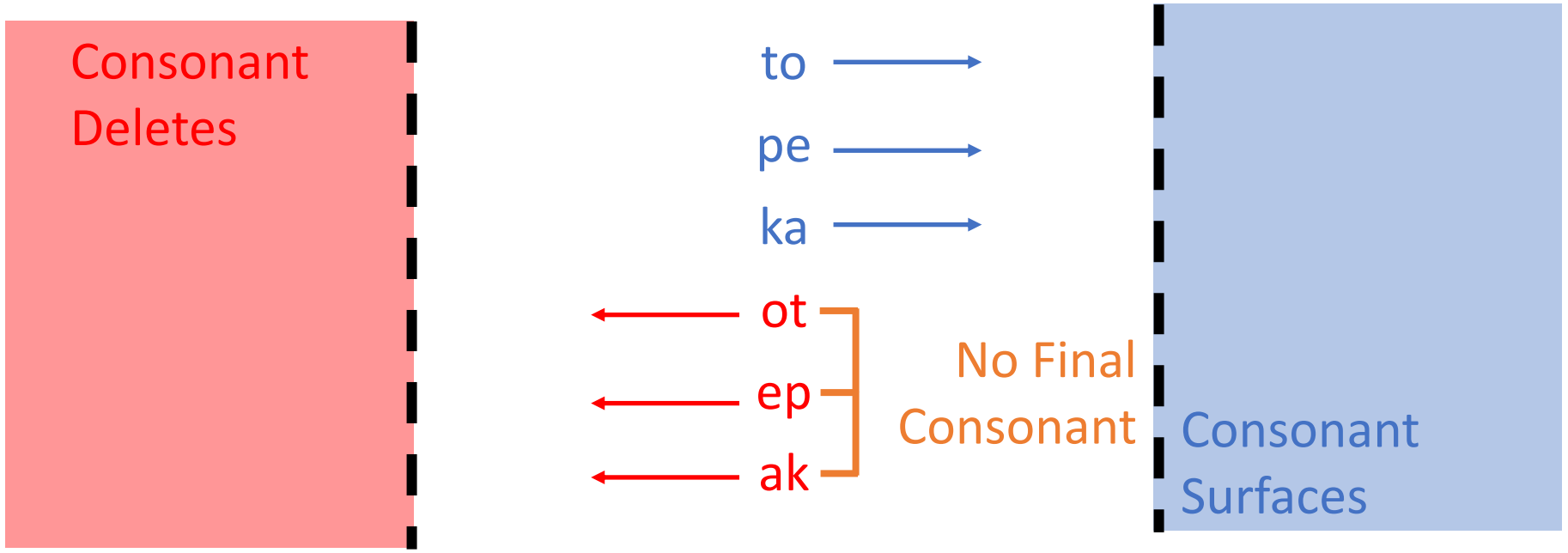
Parent Produces /ak/-[a] Child Produces /ak/-[ak]

Parent Produces /pe/-[pe] Child Produces /pe/-[e]



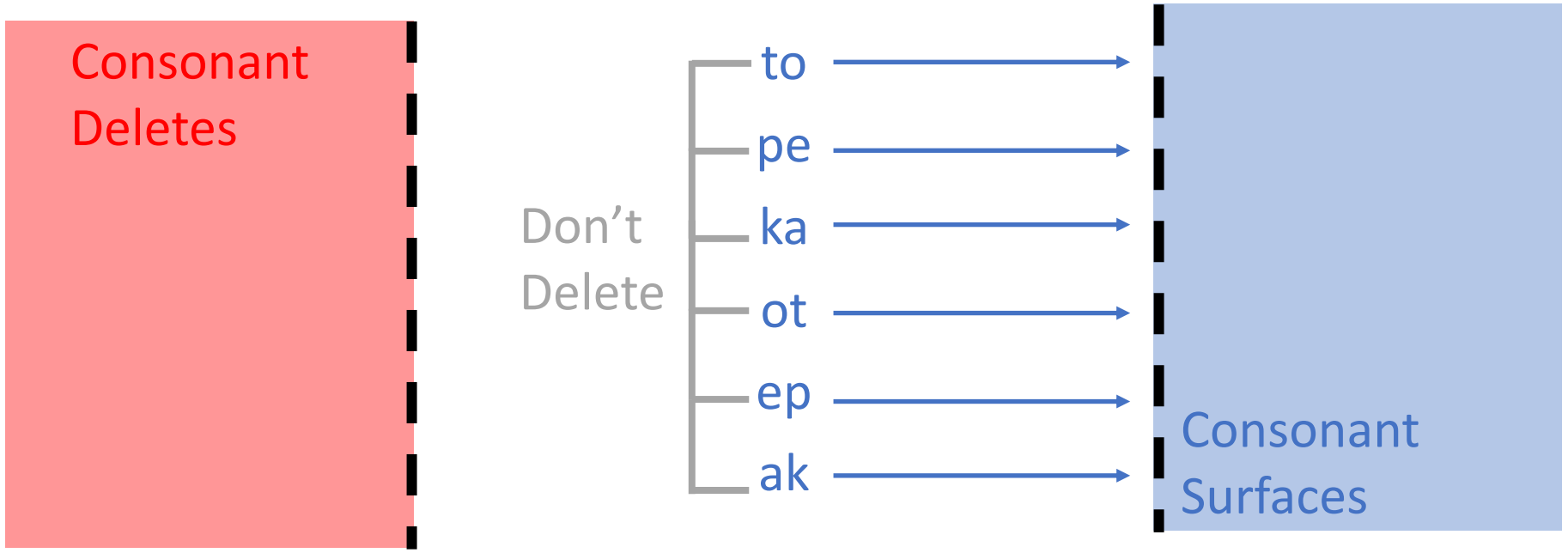
Why are some patterns more learnable?

- Average Update across possible errors.
- No form in the pattern violates **No Final Consonant.**



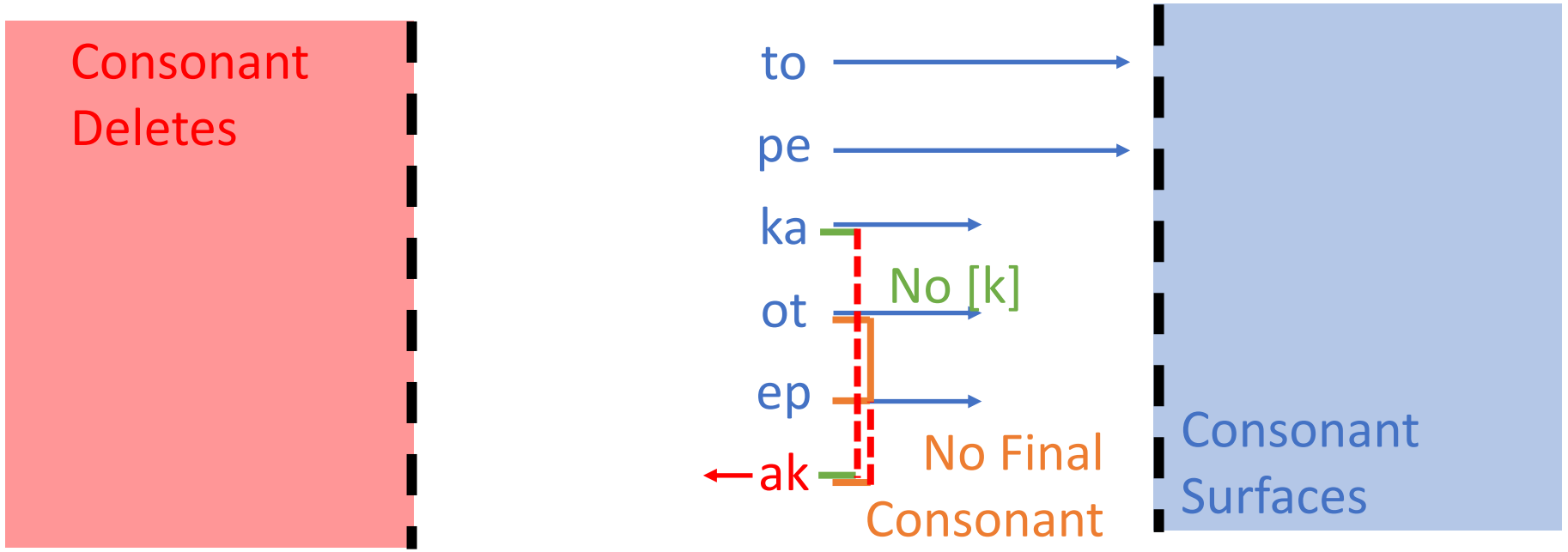
Why are some patterns more learnable?

- Average Update across possible errors.
- No form in the pattern violates **Don’t Delete.**



Why are some patterns more learnable?

- Learning a subset pattern takes longer than other patterns, because similar forms overwhelm lone dissenters.
- Target forms violate **Don't Delete, No Final Consonant,** and **No [k]**.



Takeaways

All-or-Nothing Restrictions

- More common cross-linguistically
- More stably learned across generations
- Easier to learn
- Use general constraints consistently

Subset Restrictions

- Less common cross-linguistically
- Less stably learned across generations
- Harder to learn
- Use general constraints less consistently

Lexical Frequency affects Learnability

(O’Hara 2019a, O’Hara 2020a)

Language Specific Lexical Frequency

(O'Hara 2019a, O'Hara 2020a)

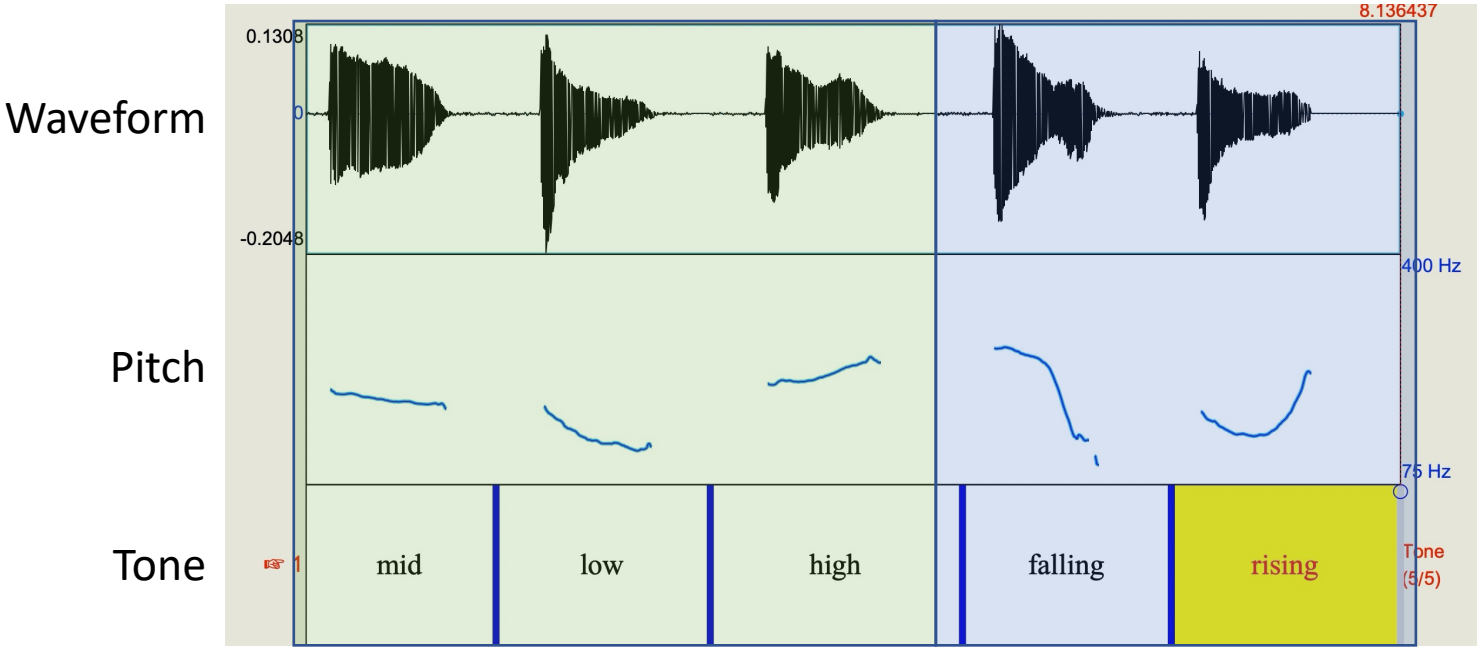
- The previous study showed how structural properties of certain patterns could affect their relative learnability in a general sense.
- Language specific lexical frequencies can also influence what patterns are easier to learn.

Lexical Frequency and Grammar

- Languages tend to have stronger restrictions in syllable types that are uncommon in their lexicon.
- This is hard to capture with the grammar.
- Learning can capture this association naturally.

Contour Tones

- In many languages, words made up of the same consonant and vowel sounds can have different meanings based on the **tone** or pitch patterns.
- Tones can be divided into **level tones** and **contour tones**



Contour Tone Distribution

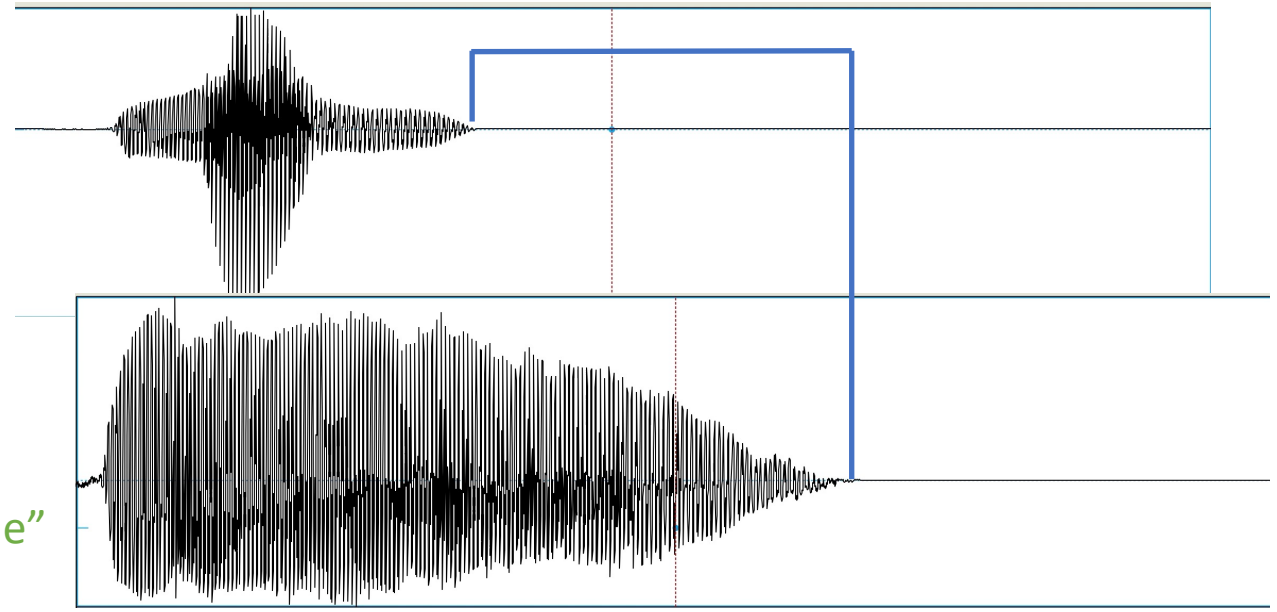
- Contour tones are more restricted than level tones.
  - Many languages allow level tones but not contour tones.
  - Many languages allow contour tones only on certain types of syllables.

Contour Tone Distribution

- Contour tones are more restricted than level tones.
  - Many languages allow level tones but not contour tones.
  - Many languages allow contour tones only on certain types of syllables.
- Contour tones are more complex, easier on higher duration syllables.
  - Better in syllables with **long vowels**, rather than syllables with **short vowels.**

Short Vowel
[nam]
“guide”

Long Vowel
[naan]
“for a long time”

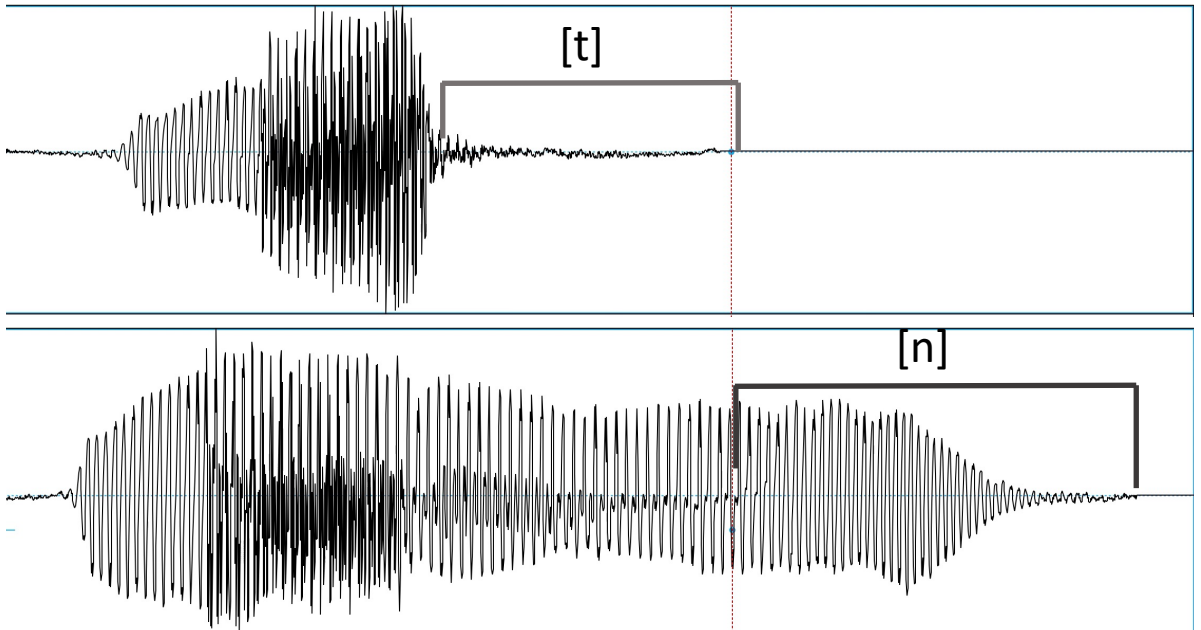


Contour Tone Distribution

- Contour tones are more restricted than level tones.
- Contour tones are more complex, easier on higher duration syllables.
  - Better in syllables with **long vowels**, rather than syllables with **short vowels.**
  - Better in **unchecked syllables** than **checked syllables.**

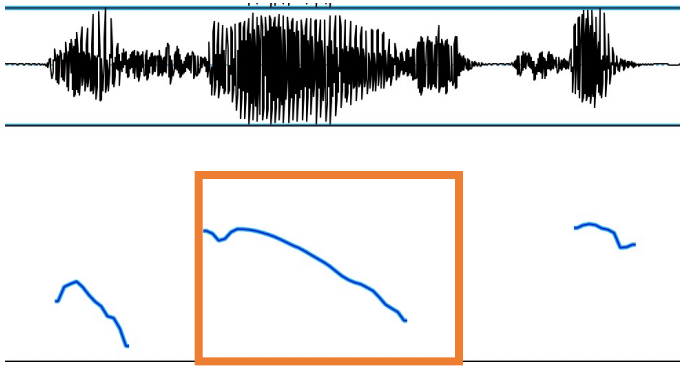
Checked Syllable
[mét]
“seed”

Unchecked Syllable
[měn]
“smelly”

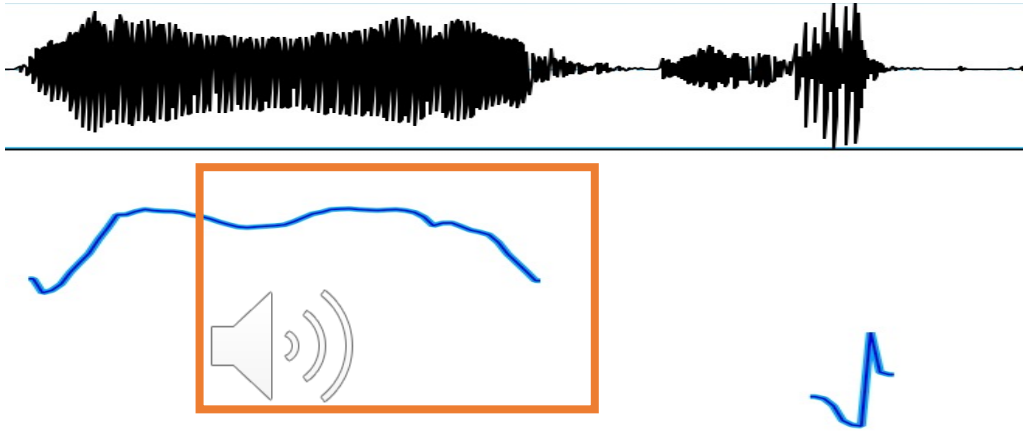


Navajo Contour Tones

- Contour Tones in Navajo are allowed in syllables with long vowels (or diphthongs), regardless of whether they are checked or unchecked.
- Checked
  - [**tě**ɪ**ʒ**ní:ɬton] `they shot at him’
  - [na**hă**:**z**tá] `they are sitting’



- Unchecked
  - [**tě**ɪ**l**ʔá] `they extend’
  - [í:**nǐl**ta] `we (2+) are studying’



Cross-linguistic Differences

Thai (and Cantonese):
Contour tones are not allowed on **checked syllables.**

	Checked	Unchecked
Short	*[**lăk**]	[**lăŋ**] `back’
Long	*[**lă:k**]	[**lă:ŋ**] `grandchild’

Navajo (and Somali):
Contour tones are not allowed on syllables with short vowels.

	Checked	Unchecked
Short	*[pìt̆ɬ]	*[pìkʰĭn]
Long	[**tĕɪʒ**ní:ɬton] `they shot at him’	[**tĕɪl**ʔá] `they extend’

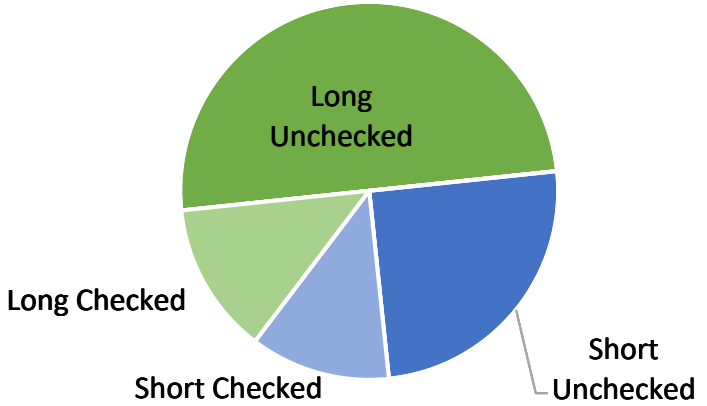
Syllable Frequency and Contour Tone Pattern

- Languages differ on whether it is worse for contour tones to be on syllables with **short vowels** or **checked syllables**.
- Claim: Languages where **short vowels** are *less common* than **checked syllables** are more likely to ban contour tones on **short vowels** than **checked syllables**.

Lexical Frequency of Syllable Types: Thai

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

	Checked	Unchecked	Total
Short	***12%***	*25%*	*37%*
Long	**13%**	50%	63%
Total	**25%**	75%	100%



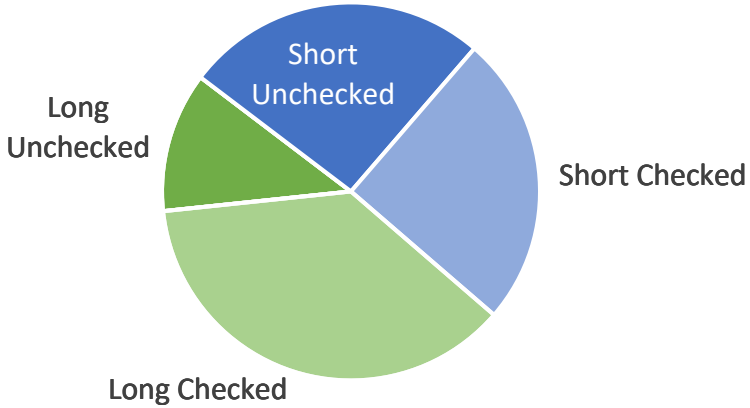
- Short syllables are more common than checked syllables

- Thai bans contour tones on checked syllables but not short syllables.

Lexical Frequency of Syllable Types: Navajo

- 39,767 words extracted from Wiktionary (Cotterell et al 2017).

	Checked	Unchecked	Total
Short	***25%***	*26%*	*51%*
Long	**37%**	12%	49%
Total	**62%**	38%	100%



- Checked syllables are more common than short syllables

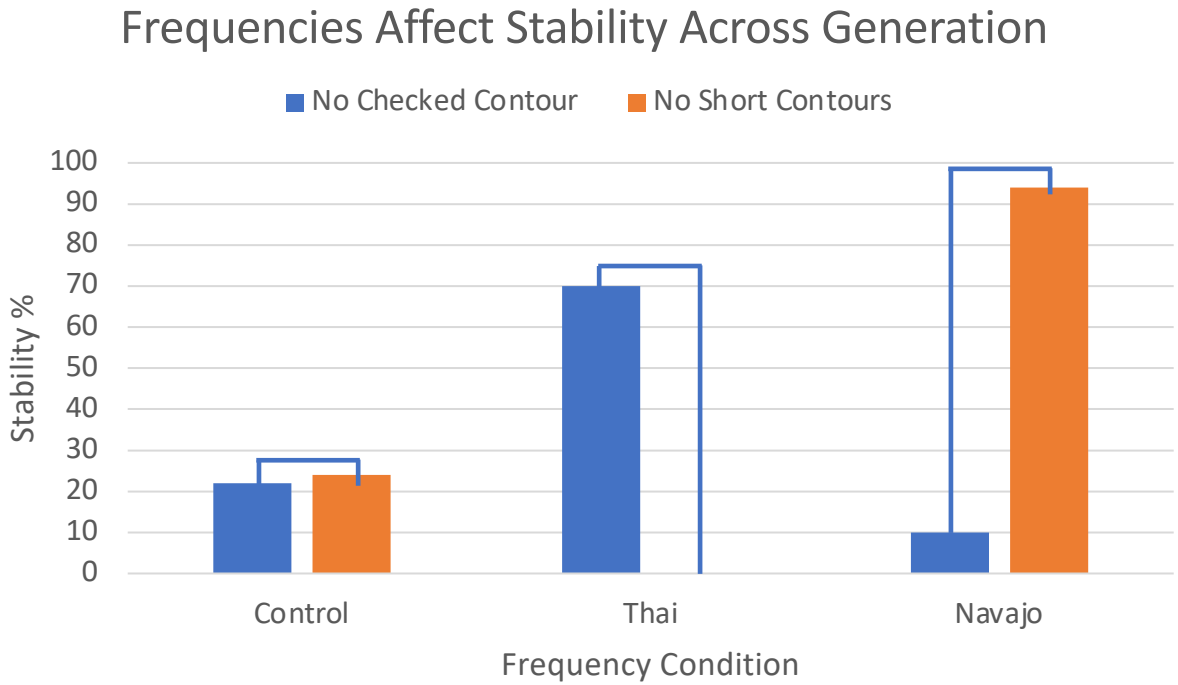
- Navajo bans contour tones on short syllables but not checked syllables.

Frequency Based Learner

	No Checked Contours		No Short Contours	
Control Frequencies Checked=Short	pat	păn	pat	pan
	paat	păan	păat	păan
Thai Frequencies Checked<Short	pat	păn	pat	pan
	paat	păan	păat	păan
Navajo Frequencies Checked>Short	pat	păn	pat	pan
	paat	păan	păat	păan

Results: Contour Tone Learning

- I ran 50 runs of each condition for 40 generations.
- With equal frequency, there is no difference in learning between the two patterns.
- With less checked syllables, like in Thai, the **No Checked Contours** pattern is easier to learn.
- With less short syllables, like in Navajo, the **No Short Contours** pattern is easier to learn.



Stability	No Checked Contours	No Short Contours
Control	24%	22%
Thai	**70%**	0%
Navajo	10%	**94%**

Takeaways

- It’s harder to learn patterns that make restrictions in common structures than rare ones
- Contour tones are lost first in less common syllable structures

- This association between frequency of syllable and amount of restriction emerges from learning.
  - Languages where common structures are more restricted are less stable.

Concluding

- Hard-to-learn patterns are less common across the world's languages.
- Learning algorithms interact with the structure of the grammar to make predictions about how common patterns should be.
- Learning allows lexical frequency to influence the grammar of a language.

Further Work

- The interaction of learning and cognitive representation offer simpler models of both.
  - Simpler cognitive frameworks (O'Hara 2019b, 2022)
  - Do constraints need to be innate? (O'Hara 2018b)
  - Simpler, more realistic Learning Algorithms (O'Hara 2017, 2020b)
- Learning allows us to disambiguate theories of mental representation.
  - Learners make use of *abstract* mental representations to learn alternations in Klamath (O'Hara 2017)
  - Neural networks emergently develop gestrual representations to handle harmony patterns (Smith et al. 2021)
  - Gestural representation accounts perform better than featural representations in the typology of harmony (Smith and O'Hara in revision)

Further Work

- Other factors that influence learnability.
  - Structural properties beyond generality (O'Hara 2021, in review).
  - Formal Language Theoretic complexity (Lamont, O'Hara, and Smith 2019, O'Hara and Smith 2019, Smith and O'Hara 2019).
  - The stability of rare and hard-to-learn patterns can be traced to rare language-specific properties (O'Hara 2018a, 2018c, 2021).

Final Word

- All languages must be learned, and transmitted across generations.
- Learners are biased towards some patterns over others.
- Through the interaction of learning and the cognitive structure of the grammar
  - We can better model more complex aspects of the asymmetries found in the world's languages
  - Develop simpler more realistic models.

Thank you!

Works Cited

- Cotterell, Ryan, Kirov, Christo, Sylak-Glassman, John, Walther, Géraldine, Vylomova, Ekaterina, Xia, Patrick, Faruqui, Manaal, and David Yarowsky, Sandra Kubler, Eisner, Jason, & Hulden, Mans. 2017. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. *Pages 1–30 of: Proceedings of the CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*.
- Dryer, Matthew S., & Haspelmath, Martin (eds). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Lamont, Andrew, O'Hara, Charlie, & Smith, Caitlin. 2019. Weakly deterministic transformations are subregular. *In: SIGMORPHON 2019: Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology and Morphology*.
- Luksaneeyanawin, S. 2000. Speech computing and speech technology in Thailand. *Interdisciplinary Approaches to Language Processing*, 267–321.
- O'Hara, Charlie. 2017. How abstract is too abstract: Learning abstract underlying representations. *Phonology*, 34(2), 325–345.
- O'Hara, Charlie. 2018a (September). *Rare Hard-To-Learn Patterns Stably Learned Due To Language-Specific Lexical Frequencies*. Talk given at Analyzing Typological Structure: From Categorical to Probabilistic Phonology. Stanford University.
- O'Hara, Charlie. 2018b (February). *Soft Typology of Coda Place of Articulation Distributions Requires Synchronic Constraints*. Talk given at the Workshop on the Emergence of Universals. Columbus OH.
- O'Hara, Charlie. 2018c (October). *The Sweet Spot Effect: Rare Phonotactic Patterns Require Specific Lexical Frequencies*. Poster presented at Annual Meeting on Phonology 2018. UCSD.
- O'Hara, Charlie. 2019a (October). *Language-Specific Factors Influence Learnability: Case Study from Contour Tone Licensing*. Poster presented at the North East Linguistic Society.
- O'Hara, Charlie. 2019b (October). *Learning Prevents MaxEnt from Giving Probability to Harmonically Bounded Candidates*. Talk given at Annual Meeting on Phonology 2019.
- O'Hara, Charlie. 2020a (January). *The Effect of Learnability on Constraint Weighting: Case Study from Contour Tone Licensing*. Poster Presented at LSA Annual Meeting 2020.

Works Cited

- O'Hara, Charlie. 2020b (January). *Frequency Matching Behavior in On-line MaxEnt Learners*. Poster presented at the Society for Computation in Linguistics (SCiL 2020).
- O'Hara, Charlie. 2021. *Soft Typology in Phonology: Learnability meets grammar*. Ph.D. thesis, University of Southern California.
- O'Hara, Charlie. 2022. MaxEnt Learners are Biased Against Giving Probability to Harmonically Bounded Candidates. In the *Proceedings of the Society for Computation in Linguistics*. 5pp
- O'Hara, Charlie. In review. *Emergent Learning Bias and the Underattestation of Simple Patterns*. Ms. University of Michigan.
- O'Hara, Charlie, & Smith, Caitlin. 2019. Computational Complexity and Sour-Grapes-Like Patterns. *In: Proceedings of the Annual Meeting on Phonology 2018*.
- Smith, Caitlin, & O'Hara, Charlie. 2019. Formal Characterizations of True and False Sour Grapes. *Pages 338–341 of: Proceedings of the Society for Computation in Linguistics*, vol. 2
- Smith, Caitlin, & O'Hara, Charlie. 2021. Learnability of derivationally opaque processes in the Gestural Harmony Model. In the *Proceedings of the Society for Computation in Linguistics*.
- Smith, Caitlin, & O'Hara, Charlie. in revision. *Learnability of derivationally opaque processes in the Gestural Harmony Model*. Ms. University of Califonia, Davis and University of Michigan.
- Smith, Caitlin, Charlie O'Hara**,** Eric Rosen, and Paul Smolensky. 2021. Emergent Gestural Scores in a Recurrent Neural Network Model of Vowel Harmony. In the *Proceedings of the Society for Computation in Linguistics*. 10pp
- Zhang, Jie. 2004. The role of contrast-specific and language-specific phonetics in contour tone distribution. *Pages 157–190 of:* Hayes, Bruce, Kirchner, Robert, & Steriade, Donca (eds), *Phonetically Based Phonology*. Cambridge University Press.

Thai sound files from slice-of-thai.com
Navajo sound files from wiktionary