Learning nonlocal phonotactics in a Strictly Piecewise phonotactic model

Huteng Dai

Rutgers University

Take-home message

- I propose a probabilistic phonotactic model and learner based on Strictly Piecewise languages studied in Formal Language Theory (FLT).
- The learner successfully learns nonlocal phonotactics from both segmental and featural representations of the corpus data, and correctly predicts the acceptability of the nonce forms in Quechua.

Formal Language Theory and noisy corpus data

- There has been a gap between FLT and noisy corpus data; (Heinz & Rawski, in press; Gouskova & Gallagher, 2020)
- The computational learning theory grounded on FLT focuses on the theorem and proof of learnability instead of simulation;
- However, understanding the domain-specific, structural properties of small dataset can help us to handle large noisy dataset.

(Heinz, 2010; Jardine & Heinz, 2016; Jardine & McMullin, 2017)

What is phonotactics?

- Phonotactics: the speakers' knowledge of possible and impossible sound sequences.

legal	*brick* [brɪk]
legal	*blick* [blɪk]
illegal	**bnick* [bnɪk]

Table 1: Local phonotactics in English

(Chomsky & Halle, 1965; Gorman, 2013)

- Nonlocal phonotactics: the phonotactic knowledge of **nonadjacent** sound sequences at **arbitrary** distance.

A running example: Quechua nonlocal phonotactics

- Quechua has three types of stops: plain stop, aspirated stop [ʰ], and ejectives [’].
- Nonlocal stop-ejective and stop-aspirate pairs are illegal in Quechua.
  - stop-ejective: *kut’u, *k’ut’u, *kʰut’u;
  - stop-aspirate: *kutʰu, *k’utʰu, *kʰutʰu;
  - legal: k’utuj ‘to cut’, rit’i ‘snow’, jutʰu ‘partridge’.

(Gouskova & Gallagher, 2020)

- Nonlocal vowel height phonotactics are also attested:
  - Uvular and high vowel sequences are illegal *q…i *q…u *i…q ……
  - Mid vowels sequences are illegal *e…e *e…o *o…e …

(Wilson & Gallagher, 2018)

Questions

- Theoretical: How do speakers learn a finite phonotactic grammar that distinguish legal and illegal words from an **infinite** set of possible sound sequences?
- Practical: can we model the phonotactic learning with input from realistic corpus data?

Local *n*-grams and baseline Learner

- Local *n*-gram: contiguous sequence of *n* items;
- Previous works usually hypothesize **local** *n*-grams as the free parameters/constraints (grammar) of the phonotactic learner.

(Hayes & Wilson, 2008)

Local *n*-grams and baseline Learner

- Local *n*-gram: contiguous sequence of *n* items;
- Previous works usually hypothesize **local** *n*-grams as the free parameters/constraints (grammar) of the phonotactic learner.

(Hayes & Wilson, 2008)

- Imagine a learner only observed one word k'utuj:

n	observed local *n*-grams	unobserved local *n*-grams
2	k'u, ut, tu, uj	*uk'...
3	k'ut, utu, tuj	*tuk'...

- E.g. *tuk'u will be penalized by the bi-/trigram constraints (*uk', *tuk').

Challenge

- However, local *n*-grams fails to capture nonlocal interactions;
- Learners based on local n-grams eventually learn numerous local n-grams that approximate nonlocal phonotactics.
- E.g. *tuʎk'u requires local 4-grams *tuʎk', *tupuk'u requires local 5-grams *tupuk', …

(Hayes & Wilson, 2008; Gouskova & Gallagher, 2020)

- Any such approximation also completely misses the generalization of nonlocal interaction at arbitrary distance.

(Heinz, 2010)

Strictly Piecewise phonotactic model

Subsequences

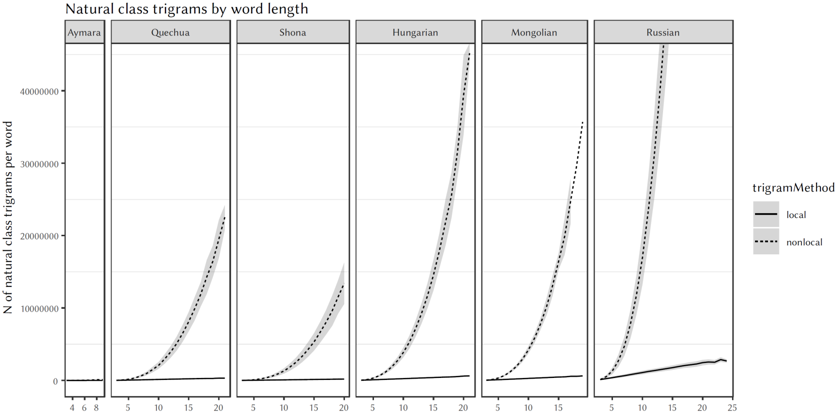
- Subsequences *aka.* nonlocal *n*-grams keep track of the order between symbols; e.g. if the learner observes k’utuj:

n	observed subsequences	unobserved subsequences
2	k’...u, k’...t , k’...j, ...	*t...k’, ...
3	k’...u...t, k’...u...j, ...	*t...u...k’, ...

- Strictly Piecewise (SP) grammar evaluates nonlocal *n*-grams; e.g. *tuk’uj, *tuʌk’u, *tupuk’u are all penalized by nonlocal bigram *t...k’ (“t precedes k’”).

(Heinz & Rogers, 2010)

Problem of exhaustively searching nonlocal *n*-grams

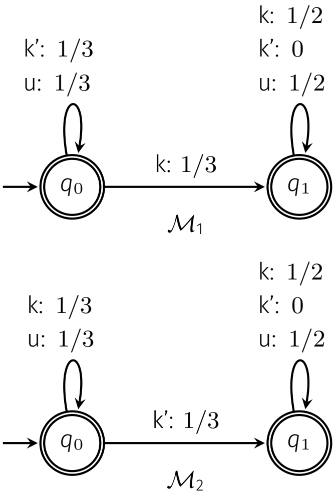


- “Devising a computationally efficient search…will require a sophisticated implementation that…is currently lacking.”

(Gouskova & Gallagher, 2020)

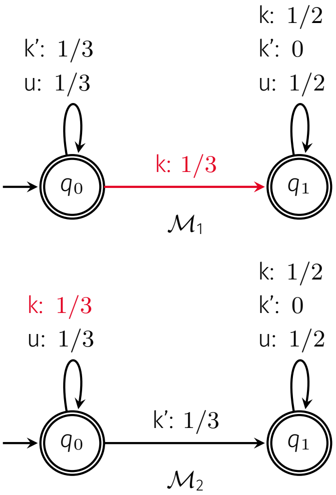
Solution: a probabilistic SP phonotactic model

- Strictly Piecewise grammar can be characterized by a set of Weighted Deterministic Finite-state Automata (WDFAs). (Shibata & Heinz, 2019)
- E.g. $\{\mathcal{M}_1, \mathcal{M}_2\}$ bans {*k…k’, *k’…k’} with a simplified alphabet A ={k, k’, u}.



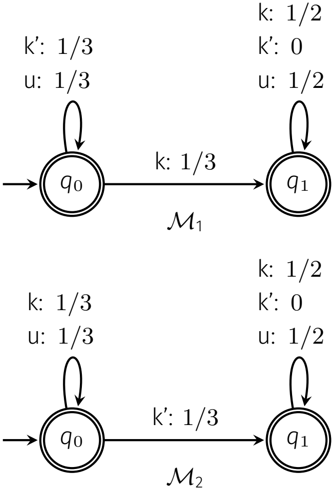
Parameters

- The parameters are transition weights $W(\mathcal{M}, q, \sigma)$ given the machine $\mathcal{M}$, state q, and segment σ.



Target symbol

- Each machine $\mathcal{M}$ only checks if it has seen one specific **target symbol** σ;
- No $\Rightarrow$ stay in state q_0;
- Yes $\Rightarrow$ go to state q_1;



Coemission probability

- Coemission probability synchronizes the parameters on different machines at the same time:

$$\text{Coemit}(\sigma_i) = \frac{\overbrace{\prod_{j=1}^{K} W(\mathcal{M}_j, q, \sigma_i)}^{\text{for one specific segment } \sigma_i}}{\underbrace{\sum_{\sigma' \in A} \prod_{j=1}^{K} W\left(\mathcal{M}_j, q, \sigma'\right)}_{\text{normalizer}}}$$

(Shibata & Heinz, 2019)

$$\begin{array}{rllllllll}
\mathcal{M}_1: & q_0 & \xrightarrow[1/3]{\text{k}} & q_1 & \xrightarrow[1/2]{\text{u}} & q_1 & \xrightarrow[0]{\text{k'}} & q_1 \\
\mathcal{M}_2: & q_0 & \xrightarrow[1/3]{\text{k}} & q_0 & \xrightarrow[1/3]{\text{u}} & q_0 & \xrightarrow[1/3]{\text{k'}} & q_1 \\
\text{Coemit}(\sigma_i): & \epsilon & \xrightarrow[1/3]{\text{k}} & \sigma_1 & \xrightarrow[1/2]{\text{u}} & \sigma_2 & \xrightarrow[0]{\text{k'}} & \sigma_3 \\
\text{Time}: & t_0 & \longrightarrow & t_1 & \longrightarrow & t_2 & \longrightarrow & t_3
\end{array}$$

Word likelihood

$$\mathcal{M}_1: \quad q_0 \xrightarrow[1/3]{\text{k}} q_1 \xrightarrow[1/2]{\text{u}} q_1 \xrightarrow[0]{\text{k'}} q_1$$

$$\mathcal{M}_2: \quad q_0 \xrightarrow[1/3]{\text{k}} q_0 \xrightarrow[1/3]{\text{u}} q_0 \xrightarrow[1/3]{\text{k'}} q_1$$

$$\text{Coemit}(\sigma_i): \epsilon \xrightarrow[1/3]{\text{k}} \sigma_1 \xrightarrow[1/2]{\text{u}} \sigma_2 \xrightarrow[0]{\text{k'}} \sigma_3$$

- Word likelihood is the product of coemission probabilities of all the segments in a word:

$$\text{lhd}(w) = \text{lhd}(\sigma_1\sigma_2\ldots\sigma_N) = \prod_{i=1}^{N} \text{Coemit}\,(\sigma_i)$$

- E.g. lhd(kuk’) $= 1/3 \cdot 1/2 \cdot 0 = 0$, Coemit(k’) $= 0$ given k $\quad\Rightarrow$ *k…k’ is penalized.

Word likelihood

$$\mathcal{M}_1: \quad q_0 \xrightarrow[1/3]{\text{k}} q_1 \xrightarrow[1/2]{\text{u}} q_1 \xrightarrow[0]{\text{k'}} q_1$$

$$\mathcal{M}_2: \quad q_0 \xrightarrow[1/3]{\text{k}} q_0 \xrightarrow[1/3]{\text{u}} q_0 \xrightarrow[1/3]{\text{k'}} q_1$$

$$\text{Coemit}(\sigma_i): \epsilon \xrightarrow[1/3]{\text{k}} \sigma_1 \xrightarrow[1/2]{\text{u}} \sigma_2 \xrightarrow[0]{\text{k'}} \sigma_3$$

- Word likelihood is the product of coemission probabilities of all the segments in a word:

$$\text{lhd}(w) = \text{lhd}(\sigma_1\sigma_2 \ldots \sigma_N) = \prod_{i=1}^{N} \text{Coemit}(\sigma_i)$$

- E.g. lhd(kuk') $= 1/3 \cdot 1/2 \cdot 0 = 0$, Coemit(k') $= 0$ given k $\quad \Rightarrow$ *k…k' is penalized.

Learning

Learning problem

- Problem: to optimize parameters $\hat{W}(\mathcal{M}, q, \sigma)$ so that the generated distribution maximally approaches the target distribution $\mathcal{D}$.
- In practice, the parameters are optimized by minimizing the **negative log likelihood** (NLL) of a sample/wordlist S drawn from $\mathcal{D}$:

$$\hat{W}(\mathcal{M}, q, \sigma) = \arg\min_{W} - \sum_{w \in S} \log \mathsf{lhd}(w).$$

$$\Uparrow$$

Maximum Likelihood Estimation $\approx$ Maximum Entropy

Training data

- $10,848$ unlabelled legal phonological words;

Training
a h i n a ʎ a m a n t a q a
t' u k u ʧ i ʃ a w a ŋ k i
qʰ e r k i ɲ ʧ o q a
...

Evaluation

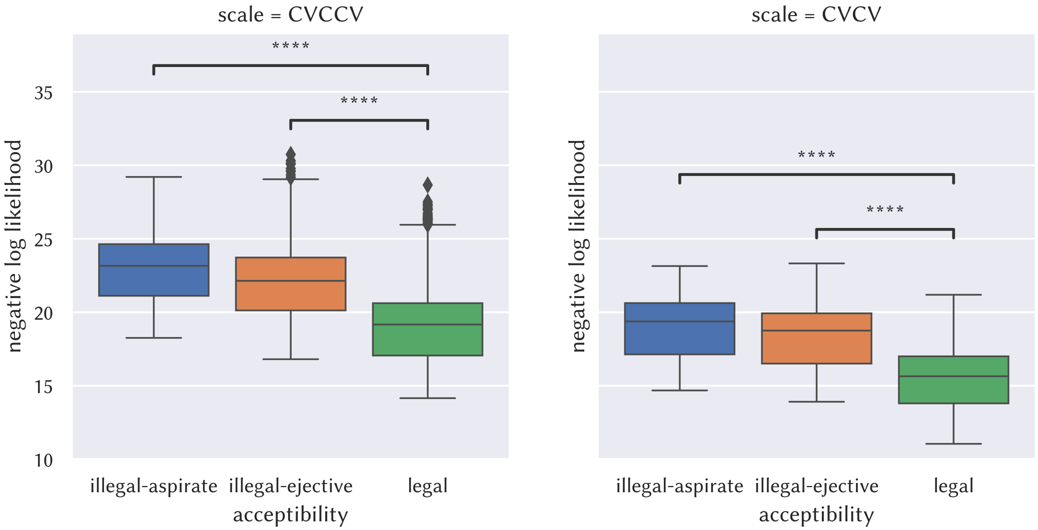
- Can’t test the accuracy since it’s unsupervised learning with unlabelled data.
- Ask if the learned model distinguish the NLL of illegal words from legal words in testing data $\rightarrow$ Clustering + nonparametric test
- If the learning is successful, legal words should have lower NLL (higher likelihood).

Testing data I: Gouskova & Gallagher (2020)

- Testing data: $24,352$ generated nonce forms (C_1VC_2V and C_1VCC_2V) which were manually labelled as legal, illegal-aspirate, and illegal-ejective. (Gouskova & Gallagher, 2020)

Testing	Label
tʃʰ a tʃʰ a	illegal-aspirate
tʃʰ a tʃʼ a	illegal-ejective
tʃʰ a ʎ tʃ a	legal
...	

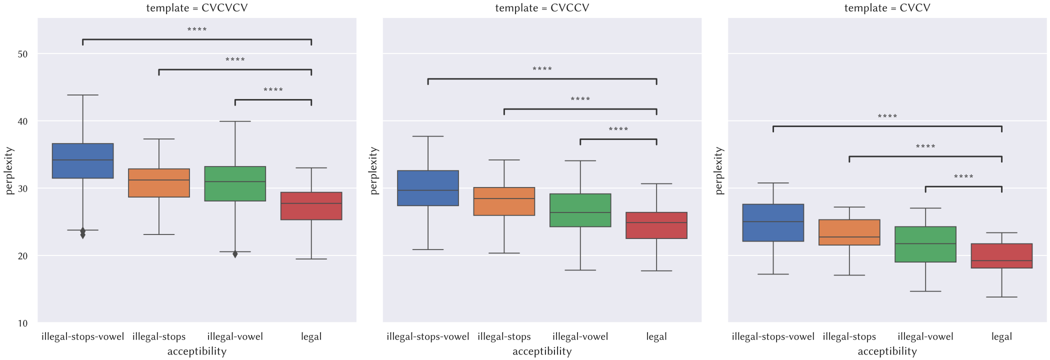
Primary result I: nonlocal phonotactics of stops



Testing data II: expanding to vowel height harmony

- Can the learned model capture the vowel height phonotactics reported in Wilson & Gallagher (2018) as well?
- 15000 generated nonce words with new labels
  - illegal-stops: violating any stop phonotactics in Gouskova & Gallagher (2020);
  - illegal-vowel: violating any vowel height phonotactics in Wilson & Gallagher (2018);
  - illegal-stops-vowel: violating any stop or vowel height phonotactics

Primary result II: interaction of multiple nonlocal phonotactics



Discussion and conclusion

Structure matters

- SP phonotactic model only keeps track of nonlocal *n*-grams, which guarantees the efficient learning of nonlocal phonotactics.
- The structure studied extensively in Formal Language Theory (FLT) is the conditions on the parameter space such as nonlocal *n*-grams.

(Heinz, 2018; Jardine & Heinz, 2016; Chandlee et al., 2019)

Acknowledgement

I thank Jeff Heinz, Adam Jardine, Bruce Tesar, Adam McCollum, Jon Rawski, and the audience at AMP 2020 for their comments and insights. My special thanks are extended to Brian Pinsky, Liam Schramm, and Yu Cao for providing the valuable suggestions on the implemented Python code.

Ineseño Chumash nonlocal sibilant phonotactics

- In Ineseño Chumash, the co-occurrence of alveolar {s, z, t͡s, d͡z,...} and lamino-postalveolar {ʃ, ʒ, t͡ʃ, d͡ʒ,...} sibilants is illegal e.g. *ʃ...s, *s...ʃ.

(Applegate, 1972)

(1) ʃapitʃʰolit /**s**-api-**tʃʰ**o-it/
'I have a stroke of good luck'

(2) ʃapitʃʰoluʃwaʃ /ʃ-api-**tʃʰ**o-u**s**-waʃ/
'He had a stroke of good luck'

(3) *sapitʃʰolit, *ʃapitʃʰoluswaʃ

3-grams	5-grams
ʃap	**ʃ**api**tʃʰ**
api	apitʃʰo
pitʃʰ	pitʃʰol
	...

Ineseño Chumash nonlocal sibilant phonotactics

- In Ineseño Chumash, the co-occurrence of alveolar {s, z, t͡s, d͡z,...} and lamino-postalveolar {ʃ, ʒ, t͡ʃ, d͡ʒ,...} sibilants is illegal e.g. *ʃ...s, *s...ʃ.

(Applegate, 1972)

(4) ʃapitʃʰolit /**s**-api-**tʃʰ**o-it/
'I have a stroke of good luck'

(5) ʃapitʃʰoluʃwaʃ /ʃ-api-**tʃʰ**o-u**s**-waʃ/
'He had a stroke of good luck'

(6) *sapitʃʰolit, *ʃapitʃʰoluswaʃ

3-grams	5-grams
ʃap	**ʃ**api**tʃʰ**
api	apitʃʰo
pitʃʰ	pitʃʰol
	...

- Trigrams won't work → difficult to choose current window *n*.

(7) ʃapitʃʰolit /**s**-api-**tʃʰ**o-it/
‘I have a stroke of good luck’

(8) ʃapitʃʰoluʃwaʃ /ʃ-api-**tʃʰ**o-u**s**-waʃ/
‘He had a stroke of good luck’

(9) *sapitʃʰolit, *ʃapitʃʰoluswaʃ

legal	illegal
ʃ...tʃʰ	*s...tʃʰ
tʃʰ...ʃ	*tʃʰ...s
ʃ...ʃ	*s...ʃ
...	...

Feature-based representation

- **Feature-based** model can be implemented by replacing the alphabet by a set of feature values $[\alpha F]$. For example, given the simple feature system below:

	F	G
a	+	-
b	+	+

Feature-based SP phonotactic model

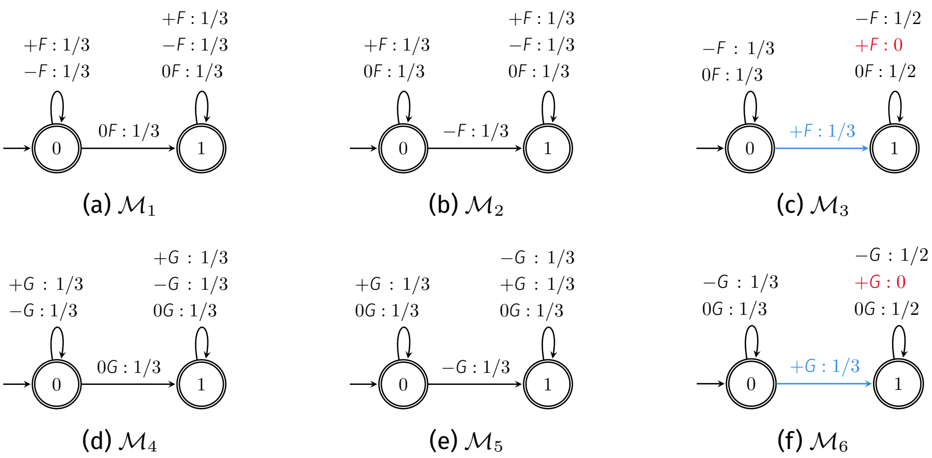
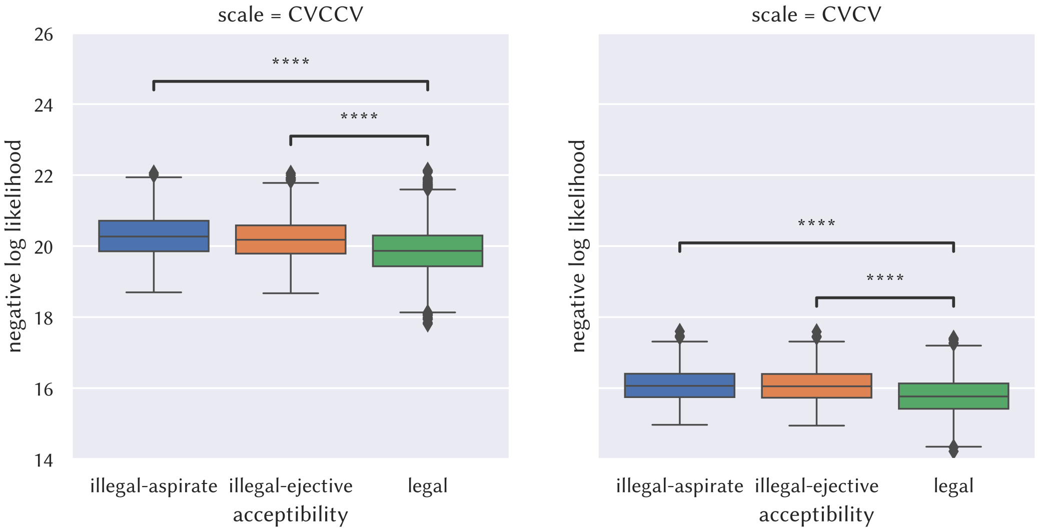


Figure 1: The feature-based SP phonotactic model which bans $^*{+F}\ldots{+F}$ and $^*{+G}\ldots{+G}$ with the simple feature system

Learning feature-based representation



Forward algorithm

NLL $\leftarrow 0$;
for *word in S* **do**
  state $\leftarrow$ 0 in each automaton $\mathcal{M}_j$;
  for σ_i *in word* **do**
    Initialize a lookup dictionary D for $\prod_{j=1}^{K} T(\mathcal{M}_j, q, \sigma')$;
    for $\mathcal{M}_j$ *in automata* **do**
      for σ' *in alphabet* **do**
        Update the lookup dictionary with σ';
      Update the state on $\mathcal{M}_j$;
    NLL $\leftarrow$ NLL - log(Coemit(σ_i))

Result: Negative log likelihood NLL of S

Reference

Applegate, R. (1972). *Ineseño chumash grammar* (Unpublished doctoral dissertation). University of California, Berkeley.

Chandlee, J., Eyraud, R., Heinz, J., Jardine, A., & Rawski, J. (2019). Learning with partially ordered representations. *arXiv preprint arXiv:1906.07886*.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of linguistics*, *1*(2), 97–138.

Gorman, K. (2013). *Generative phonotactics* (Unpublished doctoral dissertation). University of Pennsylvania.

Gouskova, M., & Gallagher, G. (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory*, 1–40.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

Heinz, J. (2010). Learning long-distance phonotactics. *Linguistic Inquiry*, *41*(4), 623–661.

Heinz, J. (2018). The computational nature of phonological generalizations. *Phonological Typology, Phonetics and Phonology*, 126–195.

Heinz, J., & Rawski, J. (in press). History of phonology: Learnability. In E. Dresher & H. van der Hulst (Eds.), *Oxford handbook of the history of phonology* (chap. 32). Oxford University Press.

Heinz, J., & Rogers, J. (2010). Estimating strictly piecewise distributions. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 886–896).

Jardine, A., & Heinz, J. (2016). Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics*, 4, 87–98.

Jardine, A., & McMullin, K. (2017). Efficient learning of tier-based strictly k-local languages. In *International conference on language and automata theory and applications* (pp. 64–76).

Shibata, C., & Heinz, J. (2019). Maximum likelihood estimation of factored regular deterministic stochastic languages. In *Proceedings of the 16th meeting on the mathematics of language (mol 16).*

Wilson, C., & Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: A case study of south bolivian quechua. *Linguistic Inquiry*, *49*(3), 610–623.