Learning Nonlocal Phonotactics in a Strictly Piecewise Probabilistic Phonotactic Model*

Huteng Dai
Rutgers University, New Brunswick

1 Introduction

Phonotactic learning is a crucial aspect of phonological acquisition and has figured significantly in computational and theoretical research in phonology. However, one persistent challenge for this line of research is inducing non-local co-occurrence patterns (Hayes & Wilson, 2008; Gouskova & Gallagher, 2020). Most previous phonotactic learners **locally** evaluates the contiguous n items (local n-grams) as phonological constraints, especially the baseline Maximum Entropy (MaxEnt) learner (Hayes & Wilson, 2008). As the length n increases, the search space grows so quickly that it becomes intractable; their learner cannot efficiently detect co-occurrence patterns over **arbitrary distances**. For instance, instead of directly penalizing the nonlocal dependency of two sibilants, the learner can only inefficiently approximate *s…ʃ by penalizing the enormous combinations of n-items e.g. trigram *soʃ, 5-gram *sopoʃ, …. Most subsequent works on MaxEnt learner generalize non-local phonotactics by searching local n-grams over postulated tiers/projections (Wilson & Gallagher, 2018). Gouskova & Gallagher (2020) further offered a method for inducing tiers from placeholder trigrams, however their learner is only shown to succeed on data in which the target phonotactics largely occur in local trigrams rather than nonlocal dependency at arbitrary distance.

The current study challenges the local n-grams as the presumed hypothesis space of phonotactic models in MaxEnt approaches, and develops a **probabilistic** phonotactic learner based on the Strictly Piecewise class of subregular languages (Heinz, 2010). The implemented learner successfully learns both segmental and featural representations of Quechua, and correctly predicts the acceptability of nonce forms in Gouskova & Gallagher (2020).

1.1 *Motivations* The current study is grounded on the "Subregular Hypothesis" (Rogers et al., 2013; Heinz, 2018) which argues that most phonological generalizations belong to a restrictive subregular region in Chomsky Hierarchy. Each subregular language can be characterized by a corresponding finite set of constraints (the *grammar*), and a Deterministic Finite-state Autotmata (de la Higuera, 2010), which can be efficiently implemented and computed. In particular, the Strictly Local (SL) and Strictly Piecewise (SP) languages correspond to the least expressive logic and lowest computational complexity.

This computational characterization of phonological typology leads to the argument that, phonotactic **learning** should search through this restrictive region to be faithful to the finite human cognition, instead of the intractable infinite hypothesis space. Specifically, previous studies have shown that Strictly Piecewise languages provide a plausible hypothesis space for non-local phonotactics (Heinz, 2010; Rogers et al., 2009; Rogers & Pullum, 2011).

There are two extremes of phonotactic models: the discrete, categorical/boolean, and qualitative one which precludes any non-categoricality in grammar, and the continuous, probabilistic, and quantitative one which denies the value of any categorical generalization (Norvig, 2012; Manning, 2003; Bod et al., 2003; Chater & Manning, 2006). Given a string (or phonological word) in a language, a categorical/boolean model predicts a binary value ("the string is/isn't in this language"), while a probabilistic model predicts a probability distribution ("the string has 1% probability of occurrence as a randomly selected word"). A probabilistic model is more favorable in handling noisy corpus data, because it assigns lower probability to,

* I thank Jeff Heinz, Adam Jardine, Bruce Tesar, Adam McCollum, Jon Rawski, Seoyoung Kim, and the audience at AMP 2020 for their valuable 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.

Proceedings of *AMP 2020*

instead of categorically penalizes, illegal forms in the corpus.

Previous studies on subregular phonology focus on the computational characterization of phonological typology instead of accounting for noisy corpus data, and usually work on non-probabilistic phonotactic models (Heinz et al., 2011; Jardine & Heinz, 2016; McMullin, 2016; Jardine & McMullin, 2017). However, it's incorrect to claim that subregular approach is incompatible with **probabilistic** phonotactic model *per se*. Following Heinz & Rogers (2010), the current study starts to bridge the gap between probabilistic approach and a subregular phonotactic model.

The current study also incorporates feature-based representation into the proposed phonotactic learner. Previous works showed that feature-based phonotactic learners are capable of handling unattested data (Albright, 2009; Wilson & Gallagher, 2018; Mayer & Nelson, 2020a). The current study focuses on implementing a feature-based phonotactic model without the full representation of natural clases such as *[+NASAL, +VOICE, ...][-NASAL, +VOICE,...] because of the unknown role of the potential feature interactions (Heinz & Koirala, 2010).

1.2 *Contributions* The current study gives insights to computational modeling of phonotactic learning by **formally** restricting the parameter space of phonotactic model. Previous works on MaxEnt model rooted in the assumption that local n-grams provide the baseline hypothesis space (Hayes & Wilson, 2008). This assumption excludes the alternative structures such as the Strictly Piecewise stringsets which naturally describe non-local dependencies (Heinz, 2007, 2010; Heinz & Idsardi, 2017).

Moreover, studying and implementing an SP phonotactic model and learner bridges the gap between the theoretical works in Formal Language Theory (FLT) and corpus data. Instead of accounting for the noisy data from natural languages, the research program of FLT has concentrated on the demarcation of linguistic typology with respect to the computational complexity. FLT approaches to learning usually assume exceptionless categorical phonotactics and symbolic phonological representation, and thus unable to handle noisy corpus data (Wilson & Gallagher, 2018; Gouskova & Gallagher, 2020).

The current study also provides tools for future study of statistical learning over other subregular classes. Although the current study focuses on SP languages, the proposed phonotactic learner can be extended to any other subregular classes, such as Strictly Local and Tier-based Strictly Local languages (Heinz et al., 2011; Jardine & Heinz, 2016; McMullin, 2016; Jardine & McMullin, 2017). There is abundant room for further progress in determining the necessary structural assumption of statistical phonotactic learning.

Furthermore, SP phonotactic model is of great theoretical interest as a variant of probabilistic finite state automata (PFA) which is formally equivalent to Hidden Markov Model (HMM) (Vidal et al., 2005a,b). SP phonotactic model is surprisingly similar to Factorial Hidden Markov Model (FHMM) (Ghahramani & Jordan, 1996; Durrieu & Thiran, 2013; Nepal & Yates, 2013) in that they both synchronize over multiple Markov chains, enabling them to make predictions based on global context. In terms of underlying structure, however, SP phonotactic model is more restrictive than FHMM because of its **deterministic** nature, and therefore closely aligns with the phonological typology argued in Subregular Hypothesis. In other words, SP phonotactic model provides a fertile ground for understanding the non-local phonological generalization with a sufficiently expressive and restrictive underlying structure.

This article is organized as follows: Section 2 introduces the Strictly Piecewise phonotactic model; Section 3 solves the learning and evaluation of SP phonotactic model; Section 4 applies the SP phonotactic model and learner to the case study of laryngeal cooccurence pattern in Quechua. A checklist of involved notations and terminologies is provided below.

2 Non-local phonotactics and Strictly Piecewise phonotactic model

Nonlocal/long-distance phonotactics is the speakers' knowledge of possible and impossible **nonadjacent** sound sequences (Gorman, 2013), which often indicates harmony patterns in input-output mappings. The current study characterizes nonlocal phonotactics by incorporating the **structure** from Strictly Piecewise grammar to generalize nonlocal phonotactics from noisy corpus data.

2.1 *SP grammar and language* A SP **grammar** evaluates **subsequences** instead of **substrings** as in n-gram models. Given a string $abcd$, the 2-long substrings include $\{ab,\ bc,\ cd\}$, while the 2-long subsequences include $\{a \dots b,\ a \dots c,\ a \dots d,\ b \dots c,\ b \dots d,\ c \dots d\}$. The length of substring/subsequences

Notations	Meanings	Examples
σ	Symbol	a, b, c
A	Alphabet	$\{a, b, c\}$
$\|A\|$	The size of the alphabet	3 for $\{a, b, c\}$
A^*	All the possible sequences with respect to A	
$\mathbb{F}$	A set of features F	$\{$VOICE, SONORANT, $\ldots\}$
$\mathbb{V}_F$	A set of feature values V_F	$\{+$VOICE, $-$SONORANT, $\ldots\}$
$\mathcal{D}$	a probabilistic distribution	
S	a finite sample of words drawn from a distribution	$\{\epsilon, abc, abccccb\}$
w	a word from a finite sample	abc
ϵ	Empty string	
$\ltimes$	Word boundary	
Pr	Probability	
lhd	Likelihood	
nll	Negative log likelihood	
$\mathcal{M}$	Finite-state machine/automata	
W	parameter weight	
Coemit	Coemission probability	

Table 1: The checklist of essential notations

is k**-factor**, which is 2 by default in the current study. The stringset generated by a SP grammar is a **SP language**. A SP grammar $\{^*a \ldots a\}$ with respect to an alphabet $A = \{a, b, c\}$ generates a SP language $\{abbbb, acccc, accbc, \ldots\}$, while bans $\{^*abbba, ^*aabbb, ^*abcba, \ldots\}$ because of the illegal subsequence $^*a \ldots a$ in the second stringset.

SP grammar precisely characterizes nonlocal phonotactics at arbitrary distance. In Ineseno Chumash, the co-occurrence of alveolar {s, z, ts, dz,… } and lamino-postalveolar {ʃ, ʒ, tʃ, dʒ, … } sibilants is illegal e.g. *ʃ…s and *z…ʒ, as illustrated in following examples:

(1) *Ineseno Chumash sibilant harmony* (Hansson, 2010; Applegate, 1972)

a. ʃapitʃʰolit /s-api-tʃʰo-it/
'I have a stroke of good luck'

b. sapitsʰolus /s-api-tʃʰo-us/
'He has a stroke of good luck'

c. ʃapitʃʰoluʃwaʃ /ʃ-api-tʃʰo-us-waʃ/
'He had a stroke of good luck'

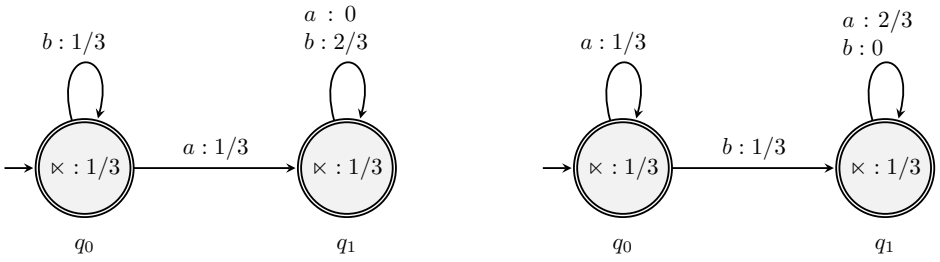
d. *sapitʃʰolit, *sapitʃʰolus, *ʃapitʃʰoluswaʃ

SP grammar characterizes the non-local phonotactics in Ineseno Chumash by banning illegal subsequences *s…tʃʰ and *tʃʰ…s.

One may convert a SP grammar to a probabilistic grammar by mapping each subsequence to real number instead of Boolean value (Heinz, 2010). Illegal subsequence ∗s…tʃʰ will be associated to lower probability e.g. 0.01, while legal subsequences s…s receives higher probability e.g. 0.99. The parameters of subsequences in a probabilistic SP grammar are similar to **violable** constraints in constraint-based grammar (Prince & Smolensky, 1993; Smolensky & Legendre, 2006). A probabilistic grammar generates a stochastic language – a probabilistic distribution over all possible strings A^*, and assigns low probability, instead of False, to a illegal string. One issue in any probabilitic grammar is that long words always receives lower probabilities (Daland, 2015), therefore the word length must be controlled in the comparison of word likelihood (see Section 4).

2.2 *SP phonotactic model and weighted automata* Gouskova & Gallagher (2020) mentioned SP language as 'nonlocal n-grams', and they claimed that it's impossible to implement a computationally

efficient search through nonlocal n-grams. The current study proposes a solution by encoding SP grammar into SP phonotactic model, which is a set of weighted deterministic finite-state automata (WDFAs) $\overrightarrow{\mathcal{M}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$. Figure 1 shows the SP phonotactic model banning $\{{}^*a \ldots a, {}^*b \ldots b\}$ ("No a following a, "No b following b") with $A = \{a, b\}$.



(a) $\mathcal{M}_1$ **(b)** $\mathcal{M}_2$

Figure 1: The SP phonotactic model banning $\{{}^*a \ldots a, {}^*b \ldots b\}$ with $A = \{a, b\}$

SP phonotactic model is **deterministic**: given a symbol in the alphabet, there is only one transition from each state. SP phonotactic model specifically characterizes SP grammar: all WDFAs in SP phonotactic model have two states, and only the transition with the target symbol enters the second state from the first state.

Each transition corresponds to certain weights which forms the **parameter** of SP phonotactic model, and the parameter weights from the second states can be interpreted as the weights of subsequences. Formally, the parameter $W(\mathcal{M}, q, \sigma) \in [0, 1]$ is the parameter weight given a factored machine $\mathcal{M}$, a segment σ, and the state q reached by its prefix.

Analysts can interpret parameters on second state q_1 in each factored machine as schematized nonlocal phonotactics, as illustrated in Table 2. y is the symbol emitting by the automata after the preceding symbol x.

$\downarrow x \rightarrow y$	a	b	$\ltimes$
a	0	2/3	1/3
b	2/3	0	1/3

Table 2: Schematized nonlocal phonotactics of Figure 1

In computing the probability of a symbol σ_i in a word w, the parameters on multiple WDFAs are synchronized by **co-emission probability**. The co-emission probability that a symbol σ_i is emitted after the SP phonotactic model reads the prefix $\sigma_1\sigma_2 \ldots \sigma_{i-1}$ is:

$$\mathsf{Coemit}(\sigma, i) = \frac{\prod_{j=1}^{K} W(\mathcal{M}_j, q, \sigma)}{\sum_{\sigma' \in A \cup \{\ltimes\}} \prod_{j=1}^{K} W(\mathcal{M}_j, q, \sigma')} \tag{1}$$

For each WDFA $\mathcal{M}_j$, q is the state that the WDFA is in after reading the prefix. The likelihood of a word w of length N is the product of co-emission probabilities given the parameters $\Theta_{\mathcal{M}}$ in factored automata $\mathcal{M}$:

$$\mathsf{lhd}(w|\Theta_{\mathcal{M}}) = \mathsf{lhd}(\sigma_1\sigma_2 \ldots \sigma_N \ltimes |\Theta_{\mathcal{M}}) = \prod_{i=1}^{N+1} \mathsf{Coemit}(\sigma, i) \tag{2}$$

Figure 2 shows the path of $ababa$ and the calculation of co-emission probability.

$$\mathcal{M}_1\colon \quad a_0 \xrightarrow[1/3]{a} a_1 \xrightarrow[2/3]{b} a_1 \xrightarrow[0]{a} a_1 \xrightarrow[2/3]{b} a_1 \xrightarrow[0]{a} a_1 \xrightarrow[1/3]{\ltimes} a_1$$

$$\mathcal{M}_2\colon \quad b_0 \xrightarrow[1/3]{a} b_0 \xrightarrow[1/3]{b} b_1 \xrightarrow[2/3]{a} b_1 \xrightarrow[0]{b} b_1 \xrightarrow[2/3]{a} b_1 \xrightarrow[1/3]{\ltimes} b_1$$

$$\textsf{Coemit}\,(\sigma_i, i)\colon \quad \lambda \xrightarrow[1/3]{\textsf{Coemit}(a,1)} \sigma_1 \xrightarrow[2/3]{\textsf{Coemit}(b,2)} \sigma_2 \xrightarrow[0]{\textsf{Coemit}(a,3)} \sigma_3 \xrightarrow[0]{\textsf{Coemit}(b,4)} \sigma_4 \xrightarrow[0]{\textsf{Coemit}(a,5)} \sigma_5 \xrightarrow[1/3]{\textsf{Coemit}(\ltimes,6)} \sigma_6$$

Figure 2: The derivation of $ababa$ in a segment-based SP model

For example, after reading the first segment a, $\mathcal{M}_1$ enters state q_1 from q_0, $\mathcal{M}_2$ is in state q_0, the co-emission probability of the second segment b in $ababa$ is:

$$\begin{aligned}&\textsf{Coemit}\,(b,2) = \\ &\frac{W(\mathcal{M}_1, a_1, b)\cdot W(\mathcal{M}_2, b_0, b)}{W(\mathcal{M}_1, a_1, a)\cdot W(\mathcal{M}_2, b_0, a) + W(\mathcal{M}_1, a_1, b)\cdot W(\mathcal{M}_2, b_0, b) + W(\mathcal{M}_1, a_1, \ltimes)\cdot W(\mathcal{M}_2, b_0, \ltimes)} \\ &= \frac{2/3\cdot 1/3}{0\cdot 1/3 + 2/3\cdot 1/3 + 1/3\cdot 1/3} = \frac{2}{3}\end{aligned} \tag{3}$$

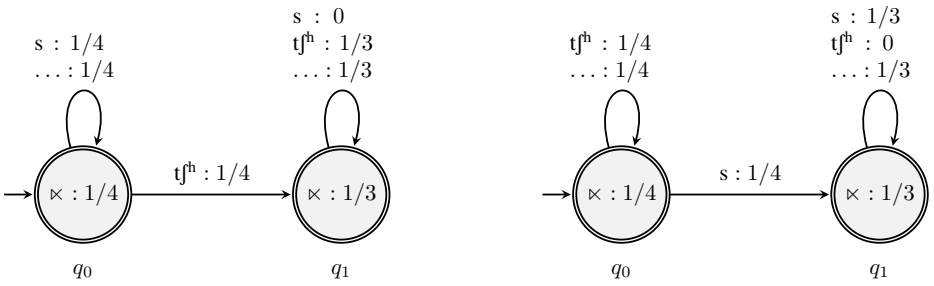
The likelihood of ab and $ababa\ltimes$ is obtained as follows:

$$\textsf{lhd}(ab \ltimes \,|\Theta_{\mathcal{M}}) = \textsf{Coemit}\,(a,1)\cdot \textsf{Coemit}\,(b,2) = 1/3\cdot 2/3 \approx 0.222$$

$$\begin{aligned}\textsf{lhd}(ababa \ltimes \,|\Theta_{\mathcal{M}}) &= \textsf{Coemit}\,(a,1)\cdot \textsf{Coemit}\,(b,2)\cdot \textsf{Coemit}\,(a,3) \\ &\quad\cdot \textsf{Coemit}\,(b,4)\cdot \textsf{Coemit}\,(a,5)\cdot \textsf{Coemit}\,(\ltimes,6) \\ &= 1/3\cdot 2/3\cdot 0\cdot 0\cdot 0\cdot 1/3 = 0\end{aligned}$$

The SP phonotactic model in Figure 1 disfavors $ababa$ than ab by assigning a lower probability to $ababa$.

SP phonotactic model can be applied to any natural languages. The model in Figure 3 models the Chumash example 1 where the probability of illegal $\textsf{Pr}(\text{s}|\text{t}\int^{\text{h}}) = \textsf{Pr}(\text{t}\int^{\text{h}}|\text{s}) = 0$ is lower than legal $\textsf{Pr}(\text{s}|\text{s}) = \textsf{Pr}(\text{t}\int^{\text{h}}|\text{t}\int^{\text{h}}) = 1/3$:



(a) $\mathcal{M}_1$ **(b)** $\mathcal{M}_2$

Figure 3: SP phonotactic model in Chumash sibilant harmony example 1; $A = \{\text{t}\int^{\text{h}}, \text{s}, \ldots\}$ where $\ldots$ indicates all the rest of the alphabet.

2.3 *Featural representation* The segment-based model above cannot capture featural phonotactics such as [+ANTERIOR]...[−ANTERIOR] and [−ANTERIOR]...[+ANTERIOR], which requires featural representation in SP phonotactic model.

The SP phonotactic model is compatible with featural representation. Formally, a **feature system** is a finite set of features $\mathbb{F} = \{F_1, \ldots, F_n\}$. Each feature is a total function which maps segments to a set

	F	G
a	+	-
b	+	+
$\ltimes$	0	0

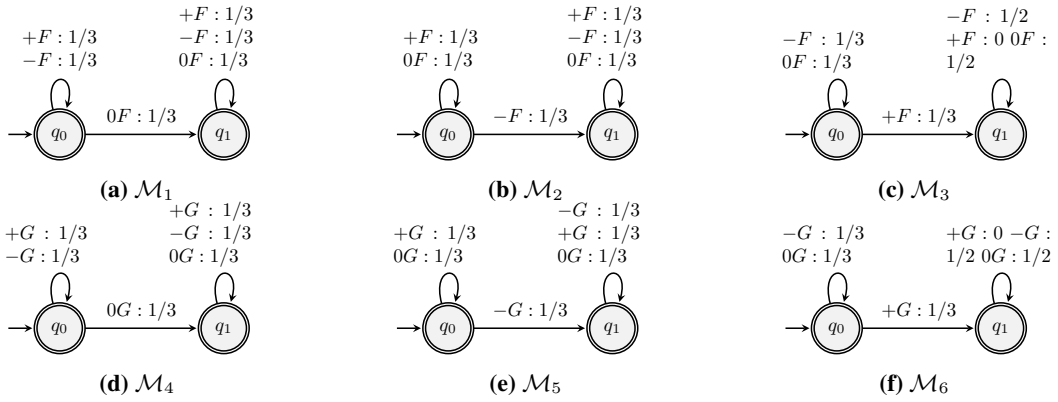
Table 3: Simple feature system with two features

of feature values $f : A \to \mathbb{V}_F$. V_F is a feature value in $\mathbb{V}_F$. For example, in the simple feature system $\mathbb{F} = \{F, G\}$ in Table 3, $\mathbb{V}_F = \{V_F : +F, -F, 0F\}$.

A feature-based SP model is the product of factored WDFAs $\mathcal{M}_j$, in which target symbols on transitions are feature values instead of segments. Figure 4 shows the feature-based SP model with respect to the simple feature system in Table 3. Each WDFA has a corresponding feature value, e.g. $\mathcal{M}_1 : 0F$, $\mathcal{M}_3 : +F$, $\mathcal{M}_4 : 0G$ ..., and three featural parameters, e.g. $\mathcal{M}_1$ corresponds to $[0F] \dots [+F]$, $[0F] \dots [-F]$, and $[0F] \dots [0F]$.

$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$		$\mathcal{M}_4$		$\mathcal{M}_5$		$\mathcal{M}_6$	
$[0F] \dots [+F]$	1/3	$[-F] \dots [+F]$	1/3	$[+F] \dots [+F]$	0	$[0G] \dots [+G]$	1/3	$[-G] \dots [+G]$	1/3	$[+G] \dots [+G]$	0
$[0F] \dots [-F]$	1/3	$[-F] \dots [-F]$	1/3	$[+F] \dots [-F]$	1/2	$[0G] \dots [-G]$	1/3	$[-G] \dots [-G]$	1/3	$[+G] \dots [-G]$	1/2
$[0F] \dots [0F]$	1/3	$[-F] \dots [0F]$	1/3	$[+F] \dots [0F]$	1/2	$[0G] \dots [0G]$	1/3	$[-G] \dots [0G]$	1/3	$[+G] \dots [0G]$	1/2

Table 4: Schematized featural nonlocal phonotactics in Figure 4



(a) $\mathcal{M}_1$ **(b)** $\mathcal{M}_2$ **(c)** $\mathcal{M}_3$

(d) $\mathcal{M}_4$ **(e)** $\mathcal{M}_5$ **(f)** $\mathcal{M}_6$

Figure 4: The 2-sets of a feature-based 2-SD-PDFA with the simple feature system which bans $+F \dots + F$ and $+G \dots + G$

The probability that a feature value $V_{F_j}(\cdot)$ for segment σ_i is emitted after the SP phonotactic model reads the prefix is:

$$\textsf{Coemit}(\sigma, i) = \frac{\prod_{j=1}^{K} W(\mathcal{M}_j, q, V_{F_j}(\sigma))}{\underbrace{\sum_{\sigma' \in A \cup \{\ltimes\}}}_{\textit{possible segments}} \prod_{j=1}^{K} W\left(\mathcal{M}_j, q, V_{F_j}(\sigma')\right)} \tag{4}$$

The model keeps track of the position of each segment, while computing the co-emission probability with respect to the feature values of each segment.

The calculation of word likelihood in feature-based SP model is the same as in segment-based SP model. The baseline feature-based model assumes the probability of one feature doesn't depend on the other feature. However, WDFAs can express parameters with certain degrees of **featural interactions** such as ${}^*[+F, -G] \dots [+F, +G]$. This issue might also be resolved by enriching the representation with natural classes, which is not treated in the current paper (see Chandlee et al. (2019) for a solution based on partially ordered structure of feature system).

3 Statistical learning in SP language model

This section addresses the learning problem in SP language model.

3.1 *Learning problem in SP language model* When the structure of WDFA $\mathcal{M}$ is known, let S be a finite sample of words drawn from the observed probabilistic distribution $\mathcal{D}$, the learning problem is to estimate the optimal parameters $\hat{\Theta}_{\mathcal{M}}$ of $\mathcal{M}$ so that the generated stochastic language maximally approaches $\mathcal{D}$. The parameters $\Theta_{\mathcal{M}}$ are parameter weights on WDFA $\mathcal{M}$.

$$\begin{aligned}\hat{\Theta}_{\mathcal{M}} &= \arg\max_{\Theta_{\mathcal{M}}}(\mathsf{lhd}(\mathcal{S}|\Theta_{\mathcal{M}})) \\ &\approx \arg\max_{\Theta_{\mathcal{M}}}(\mathsf{lhd}(D|\Theta_{\mathcal{M}}))\end{aligned} \tag{5}$$

$\mathsf{lhd}(\mathcal{S}|\Theta_{\mathcal{M}})$ is the product of probabilities for all words in the sample, which might cause underflow in practice. Instead, we transform this learning problem to log space, i.e. minimizing the negative log-likelihood of a distribution:

$$\begin{aligned}\hat{\Theta}_{\mathcal{M}} &= \arg\min_{\Theta_{\mathcal{M}}}(\mathsf{nll}(\mathcal{S}|\Theta_{\mathcal{M}})) \\ &= \arg\min_{\Theta_{\mathcal{M}}} \sum_{w \in S}(-\log \mathsf{lhd}(w|\Theta_{\mathcal{M}}))\end{aligned} \tag{6}$$

Shibata & Heinz (2019) demonstrates that this learning problem is a **convex optimization** problem (Boyd & Vandenberghe, 2004), in which the **global optimum** is guaranteed to be approximated by any algorithm.

I applied *Adam* algorithm (Kingma & Ba, 2014) to solve the optimization problem[1]. The loss function is the calculated nll, and the parameter weights which are initialized as 1. I set the learning rate to 0.005, and train the model over 20 epoches with respect to a randomized 60/40 training/validation split. In each epoch, the learner is trained on training data (60%), and the obtained model is tested on validation data (40%). The gradient of optimization is obtained with the AUTOGRAD package on PyTorch, which provides automatic differentiation for all operations in forward algorithm. Adam is applied instead of Stochastic Gradient Descent (SGD) since SGD might very quickly make those unobserved parameters 0, which might cause $\log 0$ issue.

3.2 *Evaluation* The proposed learner targets an **unsupervised learning** problem, in which only unlabelled positive evidence presented in learning data. Therefore, the learned model cannot directly predict the categorical acceptability in testing data. The learned SP model is evaluated with respect to perplexity and clustering, instead of accuracy in classification tasks. Perplexity $\rho(x)$ is the exponentiated entropy (averaged nll) of all phonemes in a dataset (Mayer & Nelson, 2020b).

Perplexity reflects the distance between the distribution predicted by a model and a testing data. The lower bound of perplexity of perplexity is 1—the closer to 1 the perplexity, the better the learned model. The upper bound of perplexity is the amount of possible random events $|x|$: each event receives an equal probability $p(x_i) = \frac{1}{|x|}$, therefore $\rho(x) = |x|$ after the derivation.

In clustering, the model assigns nll to each word in testing data in which the acceptabilities are labelled. Mann–Whitney U test (Mann & Whitney, 1947) is applied to test if the distributions of legal and illegal words are distinctive. Mann–Whitney U test is a non-parametric statistical method which counts the amount of observations from the first distribution that precede each observation from the second distribution by magnitude. A non-parametric test avoid assuming any specific shape, e.g. normal distribution, of the distributions in comparison. In the current study, the magnitude of this test is nll, which is interpreted as the grammaticality of each word. The test yields a p-value which decide whether the legal words are more likely to have a lower nll i.e. higher likelihood than illegal words.

4 Case study: Quechua

SP phonotactic model is applied to laryngeal co-occurrence patterns in (South Bolivian) Quechua.

[1] Email the author (hutengdai@gmail.com) for the Python code of the learner

4.1 *Previous work and consequence* In Quechua, nonlocal stop-ejective and stop-aspirate pairs are ill-formed ("stops" here include plain voiceless stop, ejective, and aspirated stop).

(2) *Quechua non-local restrictions on laryngeal consonants*

a. initial T' and T^h:	k'utuj	'to cut'	k^hanij	'to bite'
b. medial T' and T^h:	rit'i	'snow'	jut^{h}u	'partridge'
c. no stop . . . ejective:	*kut'u	*k'ut'u	*k^hut'u	
d. no stop . . . aspirate:	*kut^{h}u	*k'ut^{h}u	*k^hut^{h}u	

4.2 *Learning Quechua nonlocal interaction in SP phonotactic model* Both segment-based and feature-based SP phonotactic model are applied to Quechua dataset in Gouskova & Gallagher (2020), where the training data includes $10,848$ phonological words from Bolivian Quechua newspaper *Conosur Ñawpaqman*, and the testing data consists of $24,352$ nonce forms which were manually classifed as legal ($N = 18,502$), illegal-aspirate ($N = 3,645$, stop-aspirate pairs), illegal-ejective ($N = 2,205$, stop-ejective pairs). The impact of *token frequency* is not treated in the current study.

In a segment-based SP model, the perplexity is minimized to 4.8. Based on the data in Gouskova & Gallagher (2020), the baseline MaxEnt learner achieved 9.5 perplexity. Their tier-based learner, surprisingly, achieved a higher perplexity 12.9, which might suggest that their learned model is not converged, and although their tier-based learner performs well in generalizing nonlocal phonotactics, the learned distribution is in fact further away from the target distribution than the baseline learner.

The distributions of legal and illegal words are significantly distinct with respect to Mann-Whitney U test: the p-value is $2.945 \cdot 10^{-132}$ for illegal-ejective *v.s.* legal and $2.046 \cdot 10^{-185}$ for illegal-aspirate *v.s.* legal, as illustrated in Figure 5. The magnitude is negative log likelihood, and each plot includes two subplots based on the syllabic structures. The distributions are clustered with respect to three categories: illegal-asiprate, illegal-ejective, and legal.

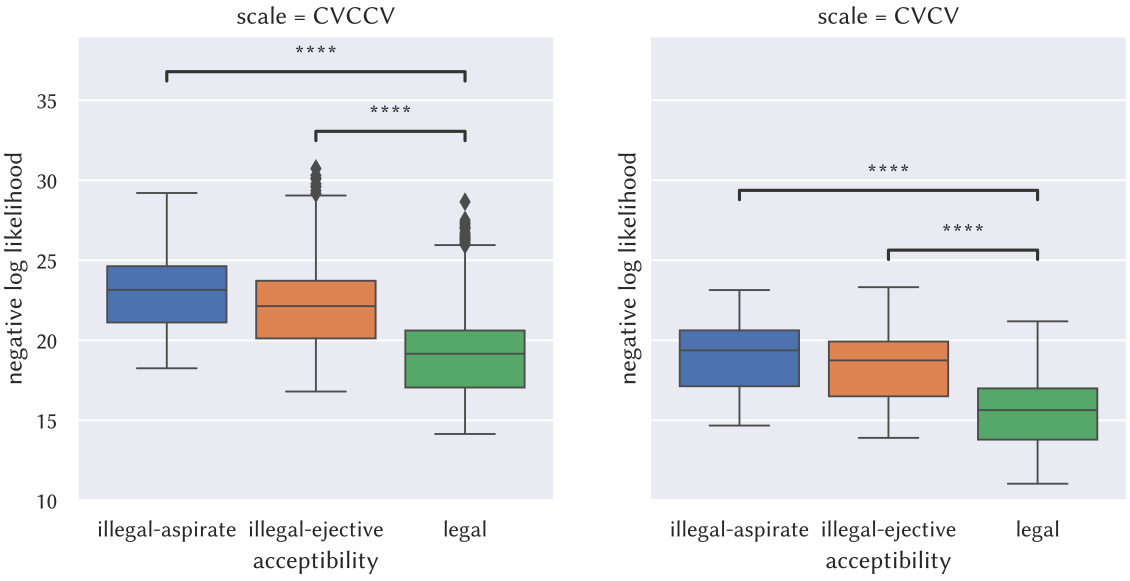


Figure 5: Boxplot negative log likelihood by segment-based SP model (Quechua; $N = 23032$)

The promising result is replicated in the feature-based model, where the model converges to 5.37 perplexity. In the clustering task, the p-value is $9.806 \cdot 10^{-37}$ for illegal-ejective *v.s.* legal and $2.113 \cdot 10^{-39}$ for illegal-aspirate *v.s.* legal, as illustrated in following boxplot:

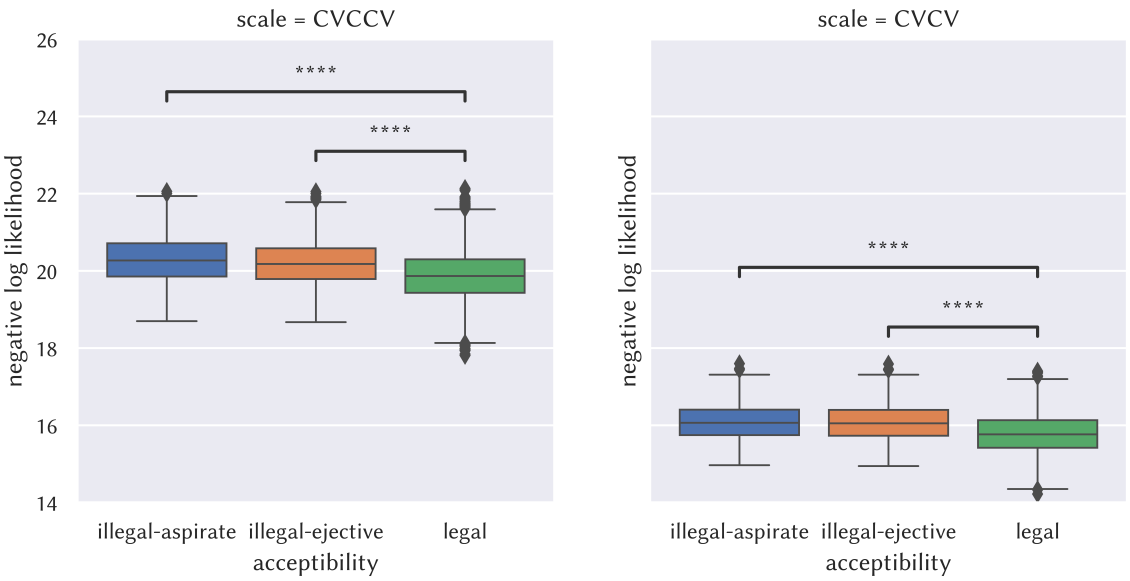


Figure 6: Boxplot negative log likelihood by feature-based SP model (Quechua; $N = 23032$)

The overlaps between the negative log likelihood of legal and illegal words are correlated with the size of parameters in a phonotactic model. That's exactly the reason why the legal and illegal words seems less distinctive in feature-based model which always has more parameters than segment-based model. Nonetheless, the statistical test justifies the argument that learnt SP model distinguishes the distribution of legal and illegal words.

To summarize, SP phonotactic learner successfully learned the model which assigns lower probability to illegal than legal words in Quechua.

5 Discussion

This section draws a comparison between SP phonotactic model and constraint-based computational models, especially Maximum Entropy (MaxEnt) model and its variants (Hayes & Wilson, 2008; Gouskova & Gallagher, 2020; Wilson & Gallagher, 2018). Moreover, the section compares subsequences to tier-based n-grams in Hayes & Wilson (2008).

5.1 *Comparison with MaxEnt approach* In learning, the current study and MaxEnt approach both follow the method of **Maximum Likelihood Estimation**, and obtain the optimal parameter weights by maximizing the likelihood of the observed forms (Mohri et al., 2018; Berger et al., 1996; Hayes & Wilson, 2008). Moreover, the implementation of MaxEnt model relies on finite-state automata as well. In Hayes & Wilson (2008), each constraint is represented as one weighted finite-state automata.

The key issue of learning phonotactics lies on the structure of a grammar, which is the abstract knowledge about the hypothesis space in learning. A reasonable and falsifiable approach relies on understanding and discovering the necessary and sufficient structure for local and nonlocal interactions. The current study and MaxEnt approaches significantly diverge in this matter.

Hayes & Wilson (2008)'s Maximum Entropy learner hypothesizes local n-grams as parameters/constraints, and cannot efficiently detect nonlocal restrictions without postulating tiers/projections. For instance, suppose a learner only recognizes one string $abcd$, the learner will hypothesize that the recognized n-grams are legal and have higher probabilities than any other possible n-grams. Meanwhile, unigrams provide the alphabet of the language.

n	legal local n-grams	illegal local n-grams
1	a, b, c, d	$*e$, $*f$, $*\ddot\smile$ …
2	ab, bc, cd	$*aa$, $*ac$, $*ad$ …
3	abc, bcd	$*aaa$, $*bbb$…

The parameter space will explode if the learner exhaustively search local n-grams to approximate nonlocal interactions. For instance, baseline MaxEnt learner will memorize local trigrams $*abc$, $*acc$, $*adc\ldots$ to approximate nonlocal constraint $*a\ldots c$. When the nonlocal phonotactics are at arbitrary distance, the hypothesis space of baseline MaxEnt learner exponentially grows ($*abba$, $*abbba$, $*abbbba$, ...), as shown in Gouskova & Gallagher (2020).

Gouskova & Gallagher (2020) induces tiers from local trigrams, instead of storing all local trigrams to approximate nonlocal interactions as in baseline MaxEnt learner. For instance, after observing $*abc$, $*acc$, $*adc$ ..., the learner will hypothesize $\{a, c\}$ as one tier. The learner will further discover tier-based local n-grams as its constraints, such as $*a\ldots c$. Constrained by the nature of local n-grams, their learner cannot induce tiers from the nonlocal interactions at **arbitrary distance**, because the learner would have to keep track of all local n-grams for any n. Gouskova & Gallagher (2020) instead proposed the heuristic that only searches local **trigrams**, as the nonlocal phonotactics in their datasets mostly exist in CVC structures. Their approach, however, cannot directly induce nonlocal interactions over more a wider window. The learner won't learn $*C_1\ldots C_2$ if the learning data only contains evidence that the constraint holds outside of a trigram window, e.g. $\mathbf{C}_1$VCCVV$\mathbf{C}_2$V.

In contrast, the structure of SP phonotactic model entails nonlocal interactions by nature as shown in the current study. This approach doesn't predict the unattested blocking effect and closely aligns with the proposal in Agreement by Correspondence (Hansson, 2010; Rose & Walker, 2004) in which subsequences, but not tier-based substrings, are the source of harmony pattern.

Another crucial difference between SP phonotactic model and MaxEnt approach is in the computation of word likelihood. MaxEnt approach assumes the word likelihood is associated to Harmony Score which is defined as the summed weights of constraint violations of a word ($\sum_i w_i \cdot C_i$), while SP phonotactic model calculates the product of co-emission probabilities of each segment. As mentioned above, the computation of Harmony score is implemented by the intersection of weighted finite state automata. However, it's an open question if harmony score can be applied to SP phonotactic model.

Besides learning and grammar, the current study also has different representational assumption comparing to the natural class-based representation in Hayes & Wilson (2008). Chandlee et al. (2019) has shown some promising result of learning natural class-based representation based on Model Theory (Libkin, 2013), while incorporating natural class-based representation to SP phonotactic model is left to future studies.

5.2 *Tier-based n-grams vs. subsequences* Most previous proposals investigating phonotactic learning have applied tiers/projections to solved the issue of nonlocal dependency. Usually the applied tiers are grouped by shared features, such as [+STRIDENT] in sibilant harmony pattern:

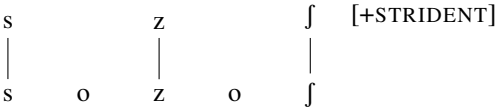


Figure 7: Tier-based representation *w.r.t.* [+STRIDENT] tier

However, these proposals also simultaneously assumes *tier-based local n-grams* (or tier-based strictly local language; TSL (Heinz et al., 2011)) as the hypothesis space, which predicts *blocking effect* in nonlocal phonotactics (Heinz, 2010). For instance, in a tier-based bi-gram model penalizes *$sʃ$ on the tier [+STRIDENT], *$soʃoz$ is penalized since s and $ʃ$ are adjacent on the tier. In contrast, $sozoʃ$ is accepted because the blocker z intervenes between s and $ʃ$. In a probabilistic model, the blocker eliminates the potential illegal substring *$sʃ$ on the tier, and $sozoʃ$ receive a higher probability than $soʃoz$. The blocking effect exists in feature-based representations as well. In Figure 8, [-ANTERIOR] of z eliminates the potential substring [-ANTERIOR][+ANTERIOR] in *$sʃ$ on the [+STRIDENT] tier.

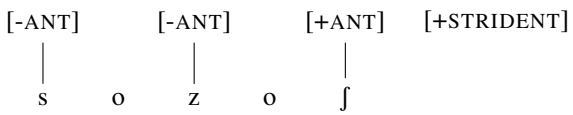


Figure 8: Tier-based featural representation of [+ANTERIOR] *w.r.t.* [+STRIDENT] tier

In contrast, searching subsequences in SP language model prevents the blocking effect. If the SP

grammar is **s…ʃ*, the *sozoʃ* and *soʃoz* are both penalized. This is true even when featural representation is entertained.

The choice between tier-based n-grams and subsequences turns out to be a typological issue. Previous studies have shown that, blocking effects are not compatible with most long-distance agreement patterns (Heinz, 2010; Rose & Walker, 2004; Hansson, 2010) with rare exceptions (McMullin, 2016). Specifically, blocking effect is not attested in the Quechua datasets. On the other hand, blocking effects are observed in some long-distance disagreement patterns, such as Latin liquid dissimilation (McMullin, 2016).

To summarize, SP language model can capture long-distance agreement patterns without the additional tier structure, and this appears to make the correct predictions in both Quechua data and typology of assimilatory harmony systems. Future research needs to examine more closely the other patterns such as disharmony with both SP and TSL language model.

6 Conclusion

The current study has proposed a probabilistic SP phonotactic model and a learning algorithm. Through a case study of Quechua laryngeal cooccurence pattern, this paper shows that SP phonotactic model precisely characterizes nonlocal phonotactics and the proposed learner generalizes both segmental and featural representations from noisy corpus data. Inspired by FHMM (Ghahramani & Jordan, 1996; Durrieu & Thiran, 2013; Nepal & Yates, 2013) and the state-of-the-art optimization algorithm (Kingma & Ba, 2014), the implementation of SP phonotactic model and learner bridges the gap between theoretical FLT approach and statistical learning, which can be further applied to other datasets from natural languages.

This paper also draws a comparison between the structural assumptions of local n-grams in MaxEnt approaches and nonlocal n-grams, or Strictly Piecewise language, in SP grammar. The current study sheds light on the scientific study of the necessary and sufficient structure for learning both local and nonlocal phonotactics.

There are several future directions. First of all, it's an open question whether SP phonotactic model can be incorporated into a constraint-based grammar by modifying the computation of word likelihood as the intersection of weighted finite state automata (Hayes & Wilson, 2008). Respectively, future work is required in MaxEnt approach to compute harmony score by means of coemisson probability as in SP phonotactic model, which naturally implicates nonlocal interaction. Another possible area of future research would be to extend the proposed phonotactic model and learner to other subregular languages, such as Strictly Local and Multi-tier Based Strictly Local languages (Heinz, 2018; Lambert & Rogers, 2020). The learning problem for any subregular languages is the same as SP language ($\hat{\Theta} = \arg\min_{\Theta_{\mathcal{M}}}(\mathsf{nll}(\mathcal{S}|\Theta_{\mathcal{M}}))$). Moreover, Shibata & Heinz (2019) has shown the *convexity* of this learning problem as long as the production of word likelihood of specific subregular language is defined with respect to coemission probability. One can easily modify each factored WDFA in SP phonotactic model to represent one tier, and model multi-tier interactions through co-emission probability. Moreover, the current study can be extended to modeling input-output phonological maps as Probabilistic Finite-state Transducers (Vidal et al., 2005b).

References

Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology* 26:1, 9–41.

Applegate, Richard (1972). *Ineseño Chumash grammar*. Ph.D. thesis, University of California, Berkeley.

Berger, Adam L, Vincent J Della Pietra & Stephen A Della Pietra (1996). A maximum entropy approach to natural language processing. *Computational linguistics* 22:1, 39–71.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (2003). *Probabilistic linguistics*. MIT Press.

Boyd, Stephen P & Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Chandlee, Jane, Remi Eyraud, Jeffrey Heinz, Adam Jardine & Jonathan Rawski (2019). Learning with partially ordered representations. *arXiv preprint arXiv:1906.07886* .

Chater, Nick & Christopher D Manning (2006). Probabilistic models of language processing and acquisition. *Trends in cognitive sciences* 10:7, 335–344.

Daland, Robert (2015). Long words in maximum entropy phonotactic grammars. *Phonology* 32:3, 353–383.

de la Higuera, Colin (2010). *Grammatical inference: learning automata and grammars*. Cambridge University Press.

Durrieu, Jean-Louis & Jean-Philippe Thiran (2013). Source/filter factorial hidden markov model, with application to pitch and formant tracking. *IEEE transactions on audio, speech, and language processing* 21:12, 2541–2553.

Ghahramani, Zoubin & Michael I Jordan (1996). Factorial hidden markov models. *Advances in Neural Information Processing Systems*, 472–478.
Gorman, Kyle (2013). *Generative phonotactics*. Ph.D. thesis, University of Pennsylvania.
Gouskova, Maria & Gillian Gallagher (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory* 1–40.
Hansson, Gunnar Ólafur (2010). *Consonant harmony: Long-distance interactions in phonology*, vol. 145. Univ of California Press.
Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry* 39:3, 379–440.
Heinz, Jeffrey (2007). *The inductive learning of phonotactic patterns*. Ph.D. thesis, PhD dissertation, University of California, Los Angeles.
Heinz, Jeffrey (2010). Learning long-distance phonotactics. *Linguistic Inquiry* 41:4, 623–661.
Heinz, Jeffrey (2018). The computational nature of phonological generalizations. *Phonological Typology, Phonetics and Phonology* 126–195.
Heinz, Jeffrey & William J Idsardi (2017). Computational phonology today. *Phonology* 34:2, 211–219.
Heinz, Jeffrey & Cesar Koirala (2010). Maximum likelihood estimation of feature-based distributions. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, Association for Computational Linguistics, 28–37.
Heinz, Jeffrey & James Rogers (2010). Estimating strictly piecewise distributions. *Proceedings of the 48th annual meeting of the association for computational linguistics*, 886–896.
Heinz, Jeffrey, Chetan Rawal & Herbert G Tanner (2011). Tier-based strictly local constraints for phonology. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, 58–64.
Jardine, Adam & Jeffrey Heinz (2016). Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics* 4, 87–98.
Jardine, Adam & Kevin McMullin (2017). Efficient learning of tier-based strictly k-local languages. *International Conference on Language and Automata Theory and Applications*, Springer, 64–76.
Kingma, Diederik P & Jimmy Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .
Lambert, Dakotah & James Rogers (2020). Tier-based strictly local stringsets: Perspectives from model and automata theory. *Proceedings of the Society for Computation in Linguistics* 3:1, 330–337.
Libkin, Leonid (2013). *Elements of finite model theory*. Springer Science & Business Media.
Mann, Henry B & Donald R Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* 50–60.
Manning, Christopher D. (2003). Probablistic syntax. Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, MIT Press, chap. 8.
Mayer, Connor & Max Nelson (2020a). Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics* 3.
Mayer, Connor & Max Nelson (2020b). Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics* 3:1, 149–159.
McMullin, Kevin James (2016). *Tier-based locality in long-distance phonotactics: learnability and typology*. Ph.D. thesis, University of British Columbia.
Mohri, Mehryar, Afshin Rostamizadeh & Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.
Nepal, Anjan & Alexander Yates (2013). Factorial hidden markov models for learning representations of natural language. *arXiv preprint arXiv:1312.6168* .
Norvig, Peter (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance* 9:4, 30–33.
Prince, Alan & Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.
Rogers, James & Geoffrey K Pullum (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information* 20:3, 329–342.
Rogers, James, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome & Sean Wibel (2009). On languages piecewise testable in the strict sense. *The mathematics of language*, Springer, 255–265.
Rogers, James, Jeffrey Heinz, Margaret Fero, Jeremy Hurst, Dakotah Lambert & Sean Wibel (2013). Cognitive and sub-regular complexity. *Formal grammar*, Springer, 90–108.
Rose, Sharon & Rachel Walker (2004). A typology of consonant agreement as correspondence. *Language* 475–531.
Shibata, Chihiro & Jeffrey Heinz (2019). Maximum likelihood estimation of factored regular deterministic stochastic languages. *Proceedings of the 16th Meeting on the Mathematics of Language (MoL 16)*.
Smolensky, Paul & Géraldine Legendre (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT press.
Vidal, Enrique, Franck Thollard, Colin de la Higuera, Francisco Casacuberta & Rafael C Carrasco (2005a). Probabilistic finite-state machines-part i. *IEEE transactions on pattern analysis and machine intelligence* 27:7, 1013–1025.
Vidal, Enrique, Frank Thollard, Colin de la Higuera, Francisco Casacuberta & Rafael C Carrasco (2005b). Probabilistic finite-state machines-part ii. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27:7, 1026–1039.
Wilson, Colin & Gillian Gallagher (2018). Accidental gaps and surface-based phonotactic learning: A case study of south bolivian quechua. *Linguistic Inquiry* 49:3, 610–623.