# Phonotactic learning with neural language models

**Connor Mayer**
Department of Linguistics
University of California, Los Angeles
connormayer@ucla.edu

**Max Nelson**
Department of Linguistics
University of Massachusetts, Amherst
manelson@umass.edu

## Abstract

Computational models of phonotactics share much in common with language models, which assign probabilities to sequences of words. While state of the art language models are implemented using neural networks, phonotactic models have not followed suit. We present several neural models of phonotactics, and show that they perform favorably when compared to existing models. In addition, they provide useful insights into the role of representations on phonotactic learning and generalization. This work provides a promising starting point for future modeling of human phonotactic knowledge.

## 1 Introduction and background

### 1.1 Phonotactics

Research on phonotactics deals broadly with two questions: what kinds of knowledge do speakers have about about the phonotactics of their language, and how is this knowledge acquired? (e.g., Chomsky and Halle, 1965) One important outcome of this work has been to show that phonotactic judgements are not categorical, but exhibit gradience: i.e., some possible words are better than others. For example, while /wɪs/ and /ploʊmf/ are both judged as being possible English words by speakers, the former is consistently judged to be a 'better' English word than the latter (Albright and Hayes, 2003; Albright, 2009). Phonotactic modelling studies have tried to build computational models of phonotactic knowledge that agree with gradient human phonotactic judgements. These models provide insight into the structure of phonological knowledge, which aspects of the data are considered by the learner when constructing their phonological grammar, and what biases constrain the forms these grammars may take (e.g., Hayes and Wilson, 2008; Albright, 2009; Daland et al., 2011; Futrell et al., 2017; Jarosz and Rysling, 2017).

### 1.2 Phonotactics and language modeling

The task undertaken by models of phonotactics is similar in many respects to the more general task of language modeling. A language model assigns probabilities to sequences of words, defining a probability distribution over word sequences (e.g., Jurafsky and Martin, 2008). A simple form of language modeling calculates $n$-gram probabilities based on corpus frequencies, and uses these to assign probabilities to longer sequences.

Phonotactic models, and models of related tasks such as word segmentation (e.g., Schrimpf and Jarosz, 2014), often frame the problem as one of language modeling over sounds rather than words. They attempt to assign probabilities to phoneme sequences that distinguish licit and illicit forms, correspond to gradient human judgements, or facilitate some task such as word segmentation. These models almost invariably operate on some version of $n$-grams, though they differ in whether they consider segments (e.g., Jelinek, 1999; Vitevitch and Luce, 2004; Jurafsky and Martin, 2008), phonological features (e.g., Albright, 2009), combinations of the two (e.g., Albright, 2009; Futrell et al., 2017), or larger prosodic structures (e.g., Coleman and Pierrehumbert, 1997; Yang, 2004; Swingley, 2005; Phillips and Pearl, 2015) to be the primitives from which sequences are built.

While early language models relied on the same types of variations on the $n$-gram employed by phonotactic learners, language modeling in NLP has seen a shift away from count-based, parametric $n$-gram models. Bengio et al. (2003) introduced a neural $n$-gram model which still makes predictions based on a fixed-size history window, but uses a neural network to generate the probability function from the history rather than simple

$n$-gram counts. Bengio et al. (2003) also introduced the idea of learning word embeddings while optimizing for the language modeling task: vector representations of words that are determined based on the word's distribution in the training data.

One shortcoming of $n$-gram models, neural or otherwise, is that the context window is fixed and specified by the researcher. This is particularly problematic for cases in which long-distance dependencies are numerous and can operate over arbitrary distances. To mitigate this issue, Mikolov et al. (2010) introduced Recurrent Neural Network Language Models (RNNLMs). These networks make use of recurrent connections to store information over potentially unbounded distances.[1] The idea of training recurrent networks on next element prediction dates to the introduction of RNNs in Elman (1990), where RNNs trained on next letter prediction were shown to learn simple phonotactic patterns like CV alternation.

Part of what the RNNLM learns is what information in the history should be considered when processing the current word. In this way RNNLMs trained on a language modeling objective are able to base predictions on all preceding information rather than just the previous $n$ words.

The RNNLM and its descendants, including LSTM language models (Sundermeyer et al., 2012) and deep contextual language models (Peters et al., 2018), have yielded dramatic improvements in performance on language modeling benchmarks, but have seen little application as phonotactic models until recently. Silfverberg et al. (2018) show that phoneme representations learned with neural methods developed for word embeddings (Word2Vec) cluster in ways that correspond to phonetic properties, and can be used to predict sound analogies. Mirea and Bicknell (2019), in a recent application of the language modeling objective to phonotactic learning, train LSTM language models on an English lexicon, and demonstrate the potential value of neural LMs as phonotactic learners.

### 1.3 The goals of this paper

The primary goal of this paper is to show that relatively simple neural network architectures developed for language modeling can be easily adapted to serve as phonotactic models, and that these models perform favorably when compared to existing models. In addition, we will show that the adoption of these neural models allows theoretical predictions about the role of representations in phonotactic grammars to be tested in ways that are not straightforward with existing models. We will demonstrate this on three phonemic data sets that exhibit phonotactic properties that have proven interesting or challenging for past models of phonotactics, and for phonological theory in general.

[1] Though in practice RNNs cannot capture arbitrarily long-distance dependencies (Bengio et al., 1994).

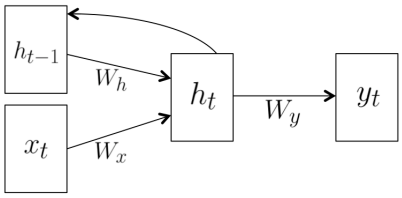


Figure 1: Schematic sRNN architecture

## 2 Model architectures

The RNNLM for phonotactic learning aims to define a probability distribution over upcoming phonemes given a representation of all preceding phonemes. We will focus on Simple Recurrent Neural Network (sRNN) variants of the models (Elman, 1990). sRNNs are a type of RNN in which the network's state at any timepoint is dependent only on the current input and the network's state at the immediately preceding timepoint (Fig. 1). The computation of the vector representing the network's state at time $t$, $h_t$, is shown in (1).

$$h_t = tanh(W_x x_t + W_h h_{t-1} + b_h) \quad (1)$$

$x_t$ is the embedding vector corresponding to the phoneme input at time $t$, $W_x$ and $W_h$ are weight matrices for the input and previous state vectors respectively, and $b_h$ is a bias vector. $h_t$ is then used to produce a probability distribution over phonemes, $\hat{y}_t$, which is the model's prediction of the identity of the segment that will appear at time $t+1$. $\hat{y}_t$ is calculated as

$$\hat{y}_t = \sigma(W_y h_t) \quad (2)$$

where $W_y$ is a weight matrix and $\sigma(z)$ is the softmax function:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (3)$$

for $i = 1, \ldots, K$.

Because the model makes predictions about upcoming data, it is able to use the same data to generate and validate its predictions, allowing unsupervised learning. At every phoneme, the cross-entropy loss is assessed between the predicted distribution before encountering that phoneme and the phoneme's one-hot encoded identity $y$:

$$L(y, \hat{y}) = -y \cdot log(\hat{y}) \quad (4)$$

All models are trained in minibatches of 64 words, which are padded to have the same length as the longest word in the batch. Loss is aggregated across each batch and backpropagated to update $W_x$, $W_h$, $W_y$, and $b_h$. Models are optimized with Adam, a variant of stochastic gradient descent that maintains individual, adaptive learning rates for all parameters (Kingma and Ba, 2014).

We build and test two distinct types of models, both of which are variants of an RNNLM, differing in their representations of phonemes. In both cases, segment identities represented by one-hot vectors are mapped to columns of an embedding weight matrix $W_E$. These vectors serve as the inputs $x_t$ for the computation in (1).

In *featural models*, the embedding vectors correspond to traditional ternary feature matrices, taken from the feature sets defined in Hayes (2009). We selected non-redundant subsets of these features for each language, and used them to construct a vector for each phoneme which specifies each feature value as positive (1), negative ($-1$), or underspecified (0). For example, the vector for English /b/ will have a 1 entry for the feature [VOICE], a $-1$ for [CONTINUANT], and a 0 for [HIGH], reflecting that [b] is a voiced non-continuant that is unspecified for height. These vectors are fixed during the learning process.

In *embedding models*, the columns of $W_E$ can take on any value in $\mathbb{R}^e$, where $e$ is a hyperparameter of the model. $W_E$ is randomly initialized and optimized alongside other model parameters, following Bengio et al. (2003). This allows the models to learn segment representations from distributional information in a way that improves performance on the language modeling objective.

Embedding models have significantly more parameters than feature models. This makes direct comparison of the two classes of models difficult, and increases the risk that embedding models overfit. To mitigate this, and to produce more interpretable embeddings, we also report results from models where the input and output embeddings are tied, following Press and Wolf (2017). The embedding weight matrix $W_E$ maps a one-hot vector of length $n$ representing a phoneme's identity to a vector of length $e$. The output weight matrix $W_y$ maps a hidden state vector $h$ to a vector of length $n$, representing a distribution over phoneme identities. Tied embeddings require that $|h| = e$, which allows for shared weights such that $W_E = W_y^T$. This functions as a kind of regularization by restricting model parameters, forcing every mapping to and from the probability distribution over phonemes to use the same set of weights.

Hyperparameter settings were chosen to optimize performance while facilitating comparison across models. Embedding models of various sizes were evaluated on a randomized 60/40 training/development split of the English data. The model that assigned the highest likelihood to the development data had 24-dimensional embeddings and 64-dimensional hidden states. These parameters were used for all embedding models. For consistency, the featural models also have 64-dimensional hidden states. Tied embedding models are trained with 24-dimensional embeddings and hidden states, ensuring a similar number of parameters to featural models. For English, there are 9,320 parameters in the embedding model, 2,248 in the featural model, and 2,200 in the tied embedding model. The number of parameters in the featural model varies slightly between languages.

The featural and embedding models instantiate different predictions about the kinds of representations used in phonotactic grammars: the featural model assumes that subsegmental representations refer only to phonetic properties, while the embedding models allow these representations to be more abstract, conditioned on how each segment patterns in the observed data. Comparison of these models allows us to computationally investigate questions that are of theoretical interest to the field, such as to what extent different types of representation help or hinder the learning of phonotactic patterns (particularly those involving phonetically unnatural classes), and the importance of representations for generalization. We return to these points in the discussion in Section 7.

## 3 Evaluation data sets

We evaluate the models on three phonotactic data sets that exhibit phenomena that have proved

challenging for previous models of phonotactics, or pose challenges for phonological theory more generally. These are Finnish vowel harmony (Section 4), Cochabamba Quechua laryngeal co-occurrence restrictions (Section 5), and English sonority projection (Section 6). Previous work suggests that models trained based on type frequency better predict human behavior than those trained on token frequency (Bybee, 1995; Albright and Hayes, 2003; Jarosz et al., 2017). We therefore do not take lexical frequency into account.

We compare the neural models against the Hayes and Wilson phonotactic learner (henceforth H&W; Hayes and Wilson, 2008). H&W is a commonly employed baseline in studies of phonotactic learning, and its use here allows the present work to be situated with respect to these studies (e.g., Albright, 2009; Daland et al., 2011; Futrell et al., 2017; Jarosz and Rysling, 2017).

H&W learns a set of featural constraints and associated weights from a training data set, and combines these constraints using a maximum entropy framework to assign probabilities to sequences of phonemes. We restrict constraint definitions to bigram or trigram windows. The Finnish and Cochabamba Quechua models learned 400 constraints, while the English model learned 600. H&W allows the analyst to specify tiers of segments over which constraints may be learned, facilitating the identification of long-distance phonotactic patterns. We compare results with and without a vowel tier for Finnish, and do not employ tiers for the other data sets.

Following Hayes and Wilson (2008), word scores for H&W are reported as maxent values ($P^*$), which for a word $x$ is calculated as

$$P^*(x) = \exp\Big(-\sum_{i=1}^{N} w_i C_i(x)\Big) \qquad (5)$$

where $N$ is the number of constraints, $w_i$ is the weight of the $i$th constraint, and $C_i(x)$ is the number of times word $x$ violates the $i$th constraint. Maxent values are proportional to probabilities: higher values indicate higher probabilities.

The RNNLM word scores are reported as perplexity ($\rho$), which is the exponentiated entropy, or inverse of the mean log likelihood, of all phonemes in the test word.

$$\rho(x) = \exp\Big(-\sum_{i=1}^{|x|} \frac{1}{|x|} log_2(p(x_i))\Big) \qquad (6)$$

| Harmonic | Disharmonic |
|---|---|
| lumo | tumæ |
| hærø | mæntu |
| mekkottastu | vastekipæ |
| pømønøritæ | testurovevy |

Table 1: Examples of harmonic and disharmonic Finnish nonce words in IPA.

Lower perplexities indicate higher probabilities.

The process of training H&W and the sRNN models is non-deterministic. H&W uses random sampling in the learning process, while the sRNN models have randomly initialized weights. We therefore report the mean scores from training and testing each model 10 times on each data set.

The model implementation and data sets are freely available online for use in future research.[2]

## 4 Finnish

### 4.1 Background

The first language we examine is Finnish. Finnish famously exhibits vowel backness harmony (e.g., Kiparsky, 1973; Ringen and Heinämäki, 1997; Goldsmith and Riggle, 2012). The language contains three classes of vowels: the front vowels {y, ø, æ}, the back vowels {u, o, a}, and the transparent vowels {i, e}. We refer to the set of front and back vowels as the harmonizing vowels. The vowels in a word generally agree in backness: that is, a word contains only transparent vowels and either front or back vowels. This restriction manifests in both root forms and affixing morphology.

This pattern is of interest because it is a long-distance phonotactic restriction. Not only can a number of consonants intervene between vowels, but an arbitrary number of transparent vowels may intervene between harmonizing vowels. This poses problems for $n$-gram models, which may not be able to detect illicit vowel subsequences if they are too far apart. We predict that the neural models will be better able to distinguish harmonic from disharmonic forms, particularly when sequences of transparent vowels occur.

### 4.2 Data

There is no publicly available corpus of transcribed Finnish. Because Finnish orthography is very close to a phonemic transcription, we instead

[2] https://github.com/MaxAndrewNelson/Phonotactic_LM

| | Harm. | Disharm. | *d* |
|---|---|---|---|
| H&W tier ($P^*$) | 0.00179 | 0.00105 | 0.46 |
| H&W no tier ($P^*$) | 0.802 | 0.708 | 0.23 |
| Feat ($\rho$) | 12.32 | 18.04 | 0.87 |
| Emb ($\rho$) | 14.97 | 25.93 | 0.86 |
| Tied Emb ($\rho$) | 11.03 | 14.42 | 0.79 |

Table 2: Average scores assigned by the models for Finnish harmonic and disharmonic words, along with effect size (Cohen's *d*).

use as training data a word list published by the Institute for the Languages of Finland.[3] We removed 584 words containing marginally attested characters, leaving 93,821 words in the corpus.

To test the models, we generated 20,000 nonce words, 10,000 harmonic and 10,000 disharmonic, ranging in length from 2–5 vowels (Table 1). Both sets are balanced for length. To ensure our models based their scores primarily on the harmony of words, we excluded CV sequences that were described to be impossible by a Finnish grammar (Suomi et al., 2008), and also excluded several CV sequences that were marginally attested in the corpus.[4] Syllables were either CV or CVC, with CC clusters drawn from the most common sequences in the corpus: /st/, /nt/, /tt/, and /kk/.

Because the test data is artificially generated, we perform no significance tests on these results. The size of the test set is arbitrary and consequently the power of the tests can be arbitrarily manipulated. Instead, we report effect sizes in the form of Cohen's $d$, which is the difference in group means expressed in units of pooled standard deviation (Cohen, 1988).

### 4.3 Results

The results are shown in Table 2. All models assign lower probabilities (lower maxent values and higher perplexities) to disharmonic forms. Cohen's $d$ indicates that the RNNLMs make this distinction more robustly: by the heuristics in Cohen (1988), the featural and embedding models display a large effect size between harmonic and disharmonic scores ($d \geq 0.8$), and the tied model displays a medium effect size ($d \geq 0.5$), while the H&W models display a small effect size ($d \geq 0.2$). Allowing H&W to use a vowel tier produces a greater distinction between harmonic and disharmonic forms, though it substantially lowers the average maxent values assigned in the test corpus.

| | Span | Harm. | Disharm. | *d* |
|---|---|---|---|---|
| H&W ($P^*$) tier | 1 | 0.00145 | 0.00131 | 0.12 |
| | 2 | 0.00138 | 0.00133 | 0.05 |
| | 3 | 0.00176 | 0.00196 | 0.16 |
| H&W ($P^*$) no tier | 1 | 0.746 | 0.707 | 0.09 |
| | 2 | 0.741 | 0.706 | 0.08 |
| | 3 | 0.804 | 0.758 | 0.13 |
| Feat ($\rho$) | 1 | 12.58 | 16.71 | 0.64 |
| | 2 | 13.10 | 16.31 | 0.38 |
| | 3 | 14.15 | 15.59 | 0.11 |
| Emb ($\rho$) | 1 | 15.79 | 21.21 | 0.57 |
| | 2 | 17.00 | 19.05 | 0.33 |
| | 3 | 16.47 | 18.94 | 0.20 |
| Tied Emb ($\rho$) | 1 | 11.49 | 13.42 | 0.61 |
| | 2 | 11.77 | 12.69 | 0.39 |
| | 3 | 11.75 | 12.61 | 0.36 |

Table 3: Model results for Finnish separated by the longest span of transparent vowels that intervene between two harmonizing vowels.

Table 3 shows that the models exhibit different performance on forms where harmonizing vowels are separated by one (e.g., [nøgihæ]; $n = 4189$), two (e.g., [jæsemehøpø]; $n = 644$), or three (e.g., [hydekistitø]; $n = 91$) transparent vowels. All models assign worse scores on average to disharmonic words, with the exception of the H&W tiered model, which assigns slightly higher scores to disharmonic words that contain spans of three transparent vowels. In addition, all models differentiate between harmonic and disharmonic forms less robustly as the maximum span of transparent vowels increases. In general, however, the RNNLMs are better able to differentiate between harmonic and disharmonic forms containing transparent vowels: the effect sizes for both H&W models on all spans is negligible ($d < 0.2$), while it is medium for all RNNLMs on spans of 1, and small on spans of 2 and 3. The exception is the featural model on spans of 3, which makes a negligible distinction. This suggests that the RNNLMs are better able to capture long distance dependencies than $n$-gram based models like H&W, even without the stipulation of a vowel tier.

## 5 Cochabamba Quechua

### 5.1 Background

The second language we examine is Cochabamba Quechua (CQ).[5] CQ has three series of stops (plain voiceless, aspirate, and ejective) at five places of articulation (labial, dental, postalveolar,

[3] http://kaino.kotus.fi/sanat/nykysuomi/

[4] These sequences are /fy/, /jø/, /fø/, /gø/, /fæ/, /gy/, /dø/, /gæ/, /bæ/, /by/, and /vø/.

[5] Thanks to Gillian Gallagher for this data.

| initial | medial | prohibited |
|---|---|---|
| t'anta | ɾit'i | *tant'a |
| k'atʃa | satʃ'a | *katʃ'a |
| p$^{h}$awaj | mosq$^{h}$oj | *posq$^{h}$oj |
| q$^{h}$ari | λimp$^{h}$i | *tʃimp$^{h}$i |

Table 4: Legal and prohibited laryngeal co-occurrence patterns in Cochabamba Quechua (Gallagher, 2019).

| | Licit | Illicit (k) | Illicit (ʁ) |
|---|---|---|---|
| H&W ($P^*$) | 0.67 | 0.28 | 0.30 |
| Feat ($\rho$) | 4.91 | 8.45 | 7.42 |
| Emb ($\rho$) | 4.89 | 8.45 | 7.55 |
| Tied Emb ($\rho$) | 4.91 | 8.28 | 7.16 |

Table 5: Model results for Cochabamba Quechua

velar, and uvular). These series participate in a laryngeal co-occurrence restriction in root forms: ejective and aspirated stops may occur either root-initially or root-medially, but they must be the first stop in the root (Table 4). Plain stops can occur following any type of stop (Gallagher, 2019).

The plain uvular stop in CQ is not realized as [q], but rather as [ʁ], a voiced uvular continuant. Gallagher (2019) provides phonetic, experimental, and phonological evidence that this phonetically disparate class (the plain stops plus [ʁ]) is active in speakers' synchronic grammars. CQ speakers preferred licit forms that do not violate the above laryngeal co-occurrence restriction to illicit forms that do, and they do not distinguish between k-initial and ʁ-initial illicit forms. For example, *[kap'a] and *[ʁap'a] are both judged as ill-formed by speakers, despite the latter appearing to satisfy the laryngeal co-occurrence restriction. Thus [ʁ] appears to pattern as a plain stop, despite being phonetically voiced and continuant.

This pattern is of interest because the set of plain stops that block the occurrence of subsequent aspirates and ejectives is a phonetically disparate class that cannot be captured with a conventional feature system, assuming [ʁ] is specified with features that reflect its phonetic realization. That is, the set of plain stops can only be specified by using disjunction between sets of features. This is primarily because [ʁ] is [+continuant], while the remaining plain stops are [–continuant]. We predict that the phonotactic models that use phonetic features may exhibit poorer performance on this pattern: specifically, we expect ʁ-initial illicit forms to receive better scores than k-initial illicit forms.

### 5.2 Data

We trained H&W and our three RNNLMs on a data set consisting of 2,468 CQ root forms. The data included two allophonic patterns related to uvular sounds: the vowels /i/ and /u/ surface as [e] and [o] respectively when adjacent to uvulars, and the sonorants /λ/, /w/, /j/, and /ɾ/ surface in uvularized forms before uvular sounds. These allophones were replaced by phonemic representations. This was done for the sake of allowing a smaller set of input segments and features to H&W, which scales poorly as the number of possible featurally-defined classes increases. This sanitization does not bear on the laryngeal co-occurrence pattern we are interested in. In addition, H&W recommends training on at least 3,000 input forms: we listed the frequency of each root as 2 in the input corpus to achieve this.

The trained models were tested on a set of 75 licit and illicit forms from Experiment 2 in Gallagher (2019). These forms were broken down into three classes: licit forms (e.g., [wap'a] or [pasi]), [k]-initial illicit forms (e.g., *[kap'a]), and [ʁ]-initial illicit forms (e.g., *[ʁap'a]). To determine whether the models assign significantly different scores to licit forms and the two types of illicit forms, we ran Kruskall-Wallis tests on each of the models with scores as the dependent variable and legality (licit vs. k-initial illicit vs. ʁ-initial illicit) as the independent variable. Kruskall-Wallis tests, which are the non-parametric equivalent of ANOVAs, were used because the scores violated several of the assumptions made by ANOVAs, such as normality of residuals. Post-hoc Dunn tests with Bonferroni correction were performed to identify significant pairwise differences.

### 5.3 Results

The results are shown in Table 5. Legality has a significant effect on score for all models (H&W: $\chi^2 = 14.53$, $p < 0.001$; Feat: $\chi^2 = 52.90$, $p < 0.001$; Emb: $\chi^2 = 53.17$, $p < 0.001$; Tied: $\chi^2 = 52.57$, $p < 0.001$). The H&W learner successfully distinguishes between licit and k-initial ($p < 0.01$) and ʁ-initial ($p < 0.05$) illicit forms, and does not make a distinction between k-initial and ʁ-initial illicit forms ($p > 0.05$). Similarly, all of the neural models are able to distinguish between licit and k-initial illicit forms (all models:

$p < 0.001$) and licit and ʁ-initial illicit forms (all models: $p < 0.001$), and not distinguish between k-initial and ʁ-initial illicit forms (all models: $p > 0.05$). Contrary to our prediction, laryngeal co-occurrence restrictions in CQ are learned by all models tested, even though this pattern makes reference to a phonetically disparate class. We can examine the models in more detail to gain insight into how this pattern is encoded in each case.

H&W cannot learn constraints that treat the plain stop series as a single class, because it cannot be uniquely specified by a feature matrix. The similar treatment of k-initial and ʁ-initial illicit forms results from multiple constraints that target different subsets of the plain stop series. For example, H&W consistently learned two high ranking constraints: *[–son, –cont]V[+CG], which penalizes illicit forms of a particular shape, except those with initial [ʁ]; and *[+dorsal, –syl]V[+CG], which penalizes only k-initial and ʁ-initial illicit forms of this shape (as well as legal but unattested forms like [xap'a]).

We may gain some insight into the neural models by comparing phoneme representations within each model using cosine similarity. Cosine similarity is the cosine of the angle between a pair of vectors: it is 1 when the vectors point in the same direction, 0 when they are orthogonal, and $-1$ when they point in opposite directions. We compare the embedding of [ʁ] with the mean of the embeddings of the classes of continuant and non-continuant consonants, which provide a representation of a 'typical' member of each class.

Table 6 shows that the representations of [ʁ] in the embedding models are more similar to the non-continuant consonants, while in the featural model it is more similar to the continuant consonants. We return to this point in the discussion.

| | continuant | non-continuant |
|---|---|---|
| Featural [ʁ] | 0.62 | 0.51 |
| Emb [ʁ] | –0.20 | 0.31 |
| Tied Emb [ʁ] | –0.26 | 0.19 |

Table 6: Cosine similarities between the embedding of [ʁ] and the mean embedding of the classes of continuant and non-continuant consonants in CQ. Learned embeddings are taken from individual runs of the models.

# 6 English

## 6.1 Background

The final phenomenon used to evaluate the neural models is English sonority projection. There is a strong preference cross-linguistically for syllables to have a sonority profile which increases monotonically from the left edge to the nucleus and then decreases from the nucleus to the right edge. This is known as the Sonority Sequencing Principle (SSP; Selkirk, 1984).

Effects of the SSP have been observed in acceptability judgments of novel words containing unattested clusters in Korean (Berent et al., 2008), Mandarin (Ren et al., 2010), English (Albright, 2007; Daland et al., 2011), and Polish (Jarosz and Rysling, 2017). The apparent universality of these effects and the fact that they apply to unattested clusters have led to a debate over whether these observations should be accounted for by an innate bias towards SSP conforming clusters (Berent et al., 2007, 2008), lexical statistics (Daland et al., 2011), or a combination of the two (Jarosz and Rysling, 2017).

We test our models on this case for two reasons. First, sonority sequencing is widely studied, particularly in English. This allows us to draw upon well-established experimental and modeling work to evaluate our results. Second, Daland et al. showed that the models that are best able to predict sonority projection from lexical statistics must have access to syllable structure and some form of subsegmental representation (for them, phonological features). Comparison of our featural and embedding models will allow us to test whether these representations must be based on phonetic properties, or if they may be learned statistically.

## 6.2 Data

All models were trained on 133,852 phonemically transcribed words in the Carnegie Mellon University Pronouncing Dictionary (CMU: Weide, 1998). Stress assignment information was removed. Words were not syllabified.

Trained models were evaluated against publicly available experimental results from Daland et al. (2011). These results come from an experiment designed to test the extent to which the sonority profile of onset clusters affects speaker acceptability judgements. Participants were tasked with choosing between pairs of nonsense words which each consisted of attested, unattested, and marginally attested English onset clusters of varying sonority profiles paired with one of six phonotactically licit tails. The onset clusters and tails

tested are shown in Table 7. The total set of words contains 96 forms: each of the 48 onsets paired with two of the tails. For each word, Daland et al. (2011) derive an aggregate goodness score. This score reflects the proportion of trials in which a word containing that cluster was chosen over its competitor.

| Onsets | | | Tails |
|---|---|---|---|
| Attested | Marginal | Unattested | |
| tw tr sw | gw ʃl | pw zr mr | -ɑtɪf |
| ʃr pr pl | vw ʃw | tl dn km | -ibɪd |
| kw kr kl | ʃn ʃm | fn ml nl | -ɑsɪp |
| gr gl fr | vl bw | dg pk lm | -ɛpɪd |
| fl dr br | dw fw | ln rl lt | -igɪf |
| bl sn sm | vr θw | rn rd rg | -ɛzɪg |

Table 7: Stimuli from Daland et al. (2011).

### 6.3 Results

Trained models were used to score the stimuli in Table 7. The success of a model was determined by the linear correlation between the mean of the model's scores across runs and the goodness scores derived from human judgements. Table 8 reports the correlation coefficients (Pearson's $r$). Following Daland et al. (2011), we report separate coefficients for words containing attested, unattested, and marginal onset clusters, as well as global correlation coefficients. The maxent values produced by H&W are positively correlated with probability, while the perplexities produced by the neural models are inversely proportional to probability. We therefore present correlations as absolute values for the sake of readability.

| | Overall | Attested | Unattested | Marginal |
|---|---|---|---|---|
| H&W (H) | 0.759 | 0.000 | 0.686 | 0.362 |
| Feat | 0.868 | 0.354 | 0.823 | 0.551 |
| Emb | 0.866 | 0.365 | 0.765 | 0.609 |
| Tied Emb | 0.853 | 0.491 | 0.738 | 0.664 |

Table 8: Correlation coefficients between model and human ratings of novel words containing attested, unattested, or marginally attested complex onsets.

All of the neural models correlate better with human judgements than H&W on every partition of the data. The high correlations between neural and human judgements across all partitions of the data demonstrate that subsegmental representations based on the phonetic properties of sounds are not necessary to effectively learn the SSP: suitable embeddings can also be learned solely from lexical statistics. This is in agreement with the findings of Mirea and Bicknell (2019), although they do not partition the data by onset type.

This is not to say, however, that there are no differences in performance between prespecified and learned embeddings. There is a tendency for the embedding models to fit observed clusters better (the attested and marginal partitions), while the featural model appears to generalize to unattested forms more effectively.

Because the available data from Daland et al. (2011) is aggregated, we are unable to use bootstrap methods to estimate the ceiling correlation coefficient, which would shed light on the extent to which human judgements would be expected to correlate with other human judgements.

| | Overall | Attested | Unattested | Marginal |
|---|---|---|---|---|
| H&W | 0.83 | 0.000 | 0.76 | 0.02 |

Table 9: Correlation coefficients between model and human judgements from the best performing model in Daland et al. (2011).

Neural models not only outperform our implementation of H&W, but perform comparably to Daland et al.'s best reported model result (Table 9), which used a version of H&W that was supplied with syllable structure. Overall these results suggest that neural phonotactic language models are able to predict aggregate human behavior as well or better than existing models even when provided with less structured input data, and that this performance does not crucially depend on whether subsegmental representations correspond to phonetic properties.

## 7 Discussion and conclusion

RNN language models can learn and generalize phonotactic patterns as well as or better than H&W across all cases considered here. The use of RNNs is particularly beneficial in the cases of Finnish and English. In Finnish, the ability of the RNN models to represent long distance dependencies allowed them to better generalize the harmony pattern to novel forms. In English, H&W generally assigns perfect scores to attested and (to a lesser extent) marginal forms, while the RNNLMs assign scores which better correlate with human judgements. Although prediction of human judgements is not the only goal of phonotactic model-

ing, it is an important one, and we believe these are useful improvements.

Comparing the performance of the models tested in this paper also provides predictions relevant to theories of universal vs. language-specific features (e.g., Mielke, 2008; Archangeli and Pulleyblank, 2018; Mayer and Daland, in press), and how this relates to the division of phonological labor between constraints and representations. The general success of the embedding models across tasks suggests these patterns may be effectively learned with no reference to segments' phonetic properties. However, it is also true that the models where segments were represented in terms of their phonetic properties were able to learn patterns involving a phonetically disparate class. The existence of such classes is a central motivation for theories of learned features.

H&W captures the CQ pattern by learning a set of constraints that, acting in tandem, produce the correct pattern. This is reminiscent of the phonological conspiracies raised by Kisseberth (1970), in that the homogeneous behavior of the plain stop series (including [ʁ]) emerges from the interaction of a set of apparently independent constraints, rather than a unified treatment by the grammar. The featural RNNLM also lacks a unified representation of this class, and we may assume the homogeneous behavior is generated by the processes applied to the representations (though these processes are computationally different from H&W). The embedding models, on the other hand, shift some of the work onto the representations, learning embeddings for [ʁ] that reflect distributional rather than phonetic properties.

Thus these models characterize different hypotheses about how phonetically disparate classes are distributed between representations and processes (e.g., rules or constraints) in the grammar. Although the performance of the featural and embedding models is indistinguishable for CQ, the results from English suggest that phonetic features may allow the models to generalize more effectively, at the expense of a poorer fit to observed data (see, e.g., Mitchell, 1980). We are optimistic that further modeling (perhaps combining fixed and learned embeddings) and comparison with human judgements will provide additional insight.

Another contribution of this paper is to show that sRNNs are able to learn phonotactic patterns as effectively as more complex models such as LSTMs (cf. Mirea and Bicknell, 2019). Phonotactic patterns are generally less complex than the syntactic/semantic patterns central to language modeling research (Heinz and Idsardi, 2013), and sRNNs may provide an appropriate fit to this complexity. For example, Weiss et al. (2018) demonstrate that, unlike LSTMs, sRNNs are unable to learn the $a^n b^n$ pattern, which is known to be phonotactically unattested (Eisner, 1997; Lamont, 2019). We anticipate for this reason that the use of more advanced models, such as attention-based language models (Vaswani et al., 2017), will not necessarily entail better performance on phonotactic learning and generalization.

Much work remains to be done. A concern with RNNLMs is that they are not as transparent as models like H&W, and are therefore of less theoretical value. Developing methods to gain insight into what these models have learned, such as probe or clustering tasks, is an important next step for their application to phonotactic learning. Such tasks can negate the interpretability problems associated with neural networks and allow access to what linguistic information is being encoded (e.g., Alishahi et al., 2019; Nelson and Mayer, 2019).

In particular, we have only shown that these models match human-like behavior in aggregate. It will be useful to explore how they deviate from human behavior in specific cases. We also note that the neural models we present here operate from left-to-right, and may have difficulty with regressive phonotactic patterns. Bidirectional RNNs (Schuster and Paliwal, 1997) have the potential to overcome this limitation.

The power of neural models as statistical learners provides a valuable tool for work on the learnability of linguistic phenomena by allowing us to begin determining the upper limit on what is learnable from lexical statistics alone, and how different representational assumptions guide this learning. We share Pater (2019)'s enthusiasm for the ongoing integration of neural research with linguistic theory as a supplement to more traditional methodology.

## Acknowledgements

We thank Gillian Gallagher, Bruce Hayes, Gaja Jarosz, Joe Pater, and the attendees of the UMass Sound Workshop. We also thank three anonymous reviewers for their valuable feedback and criticism. The authors are listed in alphabetical order.

## References


Adam Albright. 2007. Natural classes are not enough: Biased generalization in novel onset clusters. In 15th Manchester Phonology Meeting, Manchester, UK, pages 24–26.

Adam Albright. 2009. Feature-based generalization as a source of gradient acceptability. Phonology, 26:9–41.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. Cognition, 90:119–161.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. arXiv preprint arXiv:1904.04063.

Diana Archangeli and Douglas Pulleyblank. 2018. Phonology as an emergent system. In S.J. Hannahs and Anna R.K. Bosch, editors, The Routledge Handbook of Phonological Theory, pages 476–503. Routledge, London.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137–1155.

Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157–166.

Iris Berent, Tracy Lennertz, Jongho Jun, Miguel A. Moreno, and Paul Smolensky. 2008. Language universals in human brains. Proceedings of the National Academy of Sciences, 105:5321–5325.

Iris Berent, Donca Steriade, Tracy Lennertz, and Vered Vaknin. 2007. What we know about what we have never heard: Evidence from perceptual illusions. Cognition, 104:591–630.

Joan Bybee. 1995. Regular morphology and the lexicon. Language and Cognitive Processes, 10:425–455.

Noam Chomsky and Morris Halle. 1965. Some controversial questions in phonological theory. Journal of Linguistics, 1:97–138.

Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences, 2nd edition. Erlbaum, Hillsdale, NJ.

John Coleman and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In John Coleman, editor, Proceedings of the 3rd Meeting of the ACL Special Interest Group in Computational Phonology, pages 49–56. Association for Computational Linguistics, Somerset, NJ.

Robert Daland, Bruce Hayes, James White, Marc Garellek, Andreas Davis, and Ingrid Normann. 2011. Explaining sonority projection effects. Phonology, 28:197–234.

Jason Eisner. 1997. What constraints should OT allow? Handout (20p) for talk at the LSA Annual Meeting, Chicago, 1/4/97. (ROA-204-0797).

Jeffrey L. Elman. 1990. Finding structure in time. Cognitive Science, 14(2):179–211.

Richard Futrell, Adam Albright, Peter Graff, and Timothy J. O'Donnell. 2017. A generative model of phonotactics. Transactions of the Association for Computational Linguistics, 5:73–86.

Gillian Gallagher. 2019. Phonotactic knowledge and phonetically unnatural classes: the plain uvular in Cochabamba Quechua. Phonology, 36:37–60.

John Goldsmith and Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. Natural Language and Linguistic Theory, 30:859–896.

Bruce Hayes. 2009. Introductory Phonology. Wiley-Blackwell, Malden, MA.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry, 39(3):379–440.

Jeffrey Heinz and William Idsardi. 2013. What complexity differences reveal about domains in language. Topics in Cognitive Science, 5(1):111–131.

Gaja Jarosz, Shira Calamaro, and Jason Zentz. 2017. Input frequency and the acquisition of syllable structure in Polish. Language Acquisition, 24:361–399.

Gaja Jarosz and Amanda Rysling. 2017. Sonority sequencing in Polish: the combined roles of prior bias and experience. In Karen Jesney, Charlie O'Hara, Caitlin Smith, and Rachel Walker, editors, Supplemental Proceedings of the 2016 Annual Meeting on Phonology. Linguistic Society of America, Washington, DC.

Frederick Jelinek. 1999. Statistical methods for speech recognition. MIT Press, Cambridge, MA.

Dan Jurafsky and James Martin. 2008. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech processing, 2nd edition. Prentice-Hall, Upper Saddle River, NJ.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Paul Kiparsky. 1973. Phonological representations. In Osamu Fujimura, editor, Three Dimensions of Linguistic Theory, pages 1–136. TEC, Tokyo.

Charles W. Kisseberth. 1970. On the functional unity of phonological rules. Linguistic Inquiry, 1(3):291–306.

Andrew Lamont. 2019. Majority rule in harmonic serialism. In Katherine Hout, Anna Mai, Adam McCollum, Sharon Rose, and Matthew Zaslansky, editors, Supplemental Proceedings of the 2018 Annual Meeting on Phonology. Linguistic Society of America, Washington, DC.

Connor Mayer and Robert Daland. in press. A method for projecting features from observed sets of phonological classes. Linguistic Inquiry.

Jeff Mielke. 2008. The emergence of distinctive features. Oxford University Press, Oxford.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Proceedings of INTERSPEECH, pages 1045–1048.

Nicole Mirea and Klinton Bicknell. 2019. Using LSTMs to assess the obligatoriness of phonological distinctive features for phonotactic learning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1595–1605.

Tom M. Mitchell. 1980. The need for biases in learning generalizations. Report C BM-TR-5-110. New Brunswick, NJ: Rutgers University, Department of Computer Science.

Max Nelson and Connor Mayer. 2019. Learning and generalizing phonotactics with recurrent neural networks. Poster presented at the 2019 Annual Meeting on Phonology. Stonybrook, NY.

Joe Pater. 2019. Generative linguistics and neural networks at 60: foundation, friction, and fusion. Language, 93:41–74.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Lawrence Phillips and Lisa Pearl. 2015. The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. Cognitive Science, 39:1824–1854.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Jie Ren, Liqun Gao, and James L. Morgan. 2010. Mandarin speakers' knowledge of the sonority sequencing principle. In 20th Colloquium on Generative Grammar.

Catherine O. Ringen and Orvokki Heinämäki. 1997. Variation in Finnish vowel harmony: An OT account. Natural Language & Linguistic Theory, 17:303–337.

Natalie M. Schrimpf and Gaja Jarosz. 2014. Comparing models of phonotactics for word segmentation. In Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, pages 19–28. Association for Computational Linguistics, Baltimore.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45:2673–2681.

Elisabeth Selkirk. 1984. On the major class features and syllable theory. In Mark Aronoff and Richard T. Oehrle, editors, Language sound structure: Studies in phonology presented to Morris Halle by his teacher and students, pages 107–113. MIT press, Cambridge, MA.

Miikka Silfverberg, Lingshuang Jack Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. In Proceedings of the Society for Computation in Linguistics (SCiL) 2018.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association.

Kari Suomi, Juhani Toivanen, and Riikka Ylitalo. 2008. Finnish sound structure. Studia humaniora ouluensia, 9.

Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology, 50:86–132.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Michael S. Vitevitch and Paul A. Luce. 2004. A Web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, and Computers, 36:481–487.

Robert L. Weide. 1998. The CMU pronouncing dictionary. http://www.speech.cs.cmu.edu/cgibin/cmudict.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision RNNs for language recognition. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Volume 2, Short Papers, pages 740–745.

Charles D. Yang. 2004. Universal Grammar, statistics, or both. Trends in Cognitive Sciences, 8(10):451–456.