Rethinking representations: A log-bilinear model of phonotactics

Huteng Dai
Department of Linguistics
Rutgers University
huteng.dai@rutgers.edu

Connor Mayer
Department of Language Science
University of California, Irvine
cjmayer@uci.edu

Richard Futrell
Department of Language Science
University of California, Irvine
rfutrell@uci.edu

Abstract

Models of phonotactics include subsegmental representations in order to generalize to unattested sequences. These representations can be encoded in at least two ways: as discrete, phonetically-based features, or as continuous, distribution-based representations induced from the statistical patterning of sounds. Because phonological theory typically assumes that representations are discrete, past work has reduced continuous representations to discrete ones, which eliminates potentially relevant information. In this paper we present a model of phonotactics that can use continuous representations directly, and show that this approach yields competitive performance on modeling experimental judgments of English sonority sequencing. The proposed model broadens the space of possible phonotactic models by removing requirements for discrete features, and is a step towards an integrated picture of phonotactic learning based on distributional statistics and continuous representations.

1 Introduction

Phonotactics refers to restrictions on how sounds can be sequenced in a language. For example, although neither *blick* [blɪk] nor *bnick* [bnɪk] are real English words, native speakers feel that *blick* could be an English word, while *bnick* could not because it begins with the prohibited onset *[bn] (Chomsky and Halle, 1965). Phonotactic restrictions vary between languages, meaning that they must be learned. For example, *steek* [stik] is a possible word in English but not in Spanish, because the latter has a phonotactic restriction on syllables beginning with [st]. As learners acquire a language, they become sensitive to the frequencies of different sequences. This phonotactic knowledge underlies speakers' intuitions about possible words in their language.

Experimental studies involving acceptability judgments have found that speakers have **gradient intuitions** about phonotactic well-formedness (e.g., Coleman and Pierrehumbert, 1997; Albright, 2009; Hayes et al., 2009; Daland et al., 2011). For example, when considering the nonce words *blick* [blɪk], *bnick* [bnɪk], and *bwick* [bwɪk], English speakers typically find *blick* to be acceptable, *bnick* to be poor, and *bwick* to be intermediate between the two (Albright, 2009). This has led to the development of **probabilistic** models of phonotactics, which assign a continuous score to words that reflects their gradient well-formedness (Hayes and Wilson, 2008; Futrell et al., 2017; Wilson and Gallagher, 2018; Gouskova and Gallagher, 2020; Mayer and Nelson, 2020). Phonotactics is also commonly treated as probabilistic in models of higher-level linguistic tasks, such as speech perception and word segmentation (see discussion in Daland, 2015).

1.1 Feature-based generalizations

An additional difficulty for phonotactic models is the problem of **accidental gaps**: sequences of sounds that do not appear in the lexicon but are judged to be acceptable. Humans do not treat unattested sequences uniformly: in the example in the previous section, both [bw] and [bn] are unattested onsets in English, but the former is preferred to the latter. Phonotactic models thus need to be able to generalize to unseen sequences in a way that is consistent with human behavior.

The standard solution is to have models operate on **featural representations**, which decompose segments into sets of feature-value pairs (or, alternatively, a vector of values whose dimensions are the features). Features allow models to refer to classes of segments based on shared properties. In English, for example, the feature vector [−continuant] characterizes the set of stops and affricates, [−sonorant] picks out the set of obstru-

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 259-268.
Amherst, Massachusetts, June 15-17, 2023

ents, and [−continuant, −sonorant] picks out the set of obstruent stops/affricates (excluding the nasal stops). Returning to the example above, although [bw] and [bn] are both unattested onsets, there are many onsets that are featurally similar to [bw], consisting of b[+approximant] sequences like [bj], [bl], [bɹ]. There are none that are similar to [bn], consisting of b[−continuant]. Features allow these kinds of generalizations to be modeled.

1.2 Whence features?

Phonological features are typically defined with respect to phonetic properties (e.g., Chomsky and Halle, 1965). This reflects the strong typological tendency that sounds with similar phonetic properties tend to pattern similarly.

More recent research has proposed that features may be **emergent**, reflecting shared, language-specific distributional properties in addition to phonetic properties (e.g., Mielke, 2008; Archangeli and Pulleyblank, 2018; Gallagher, 2019; Archangeli and Pulleyblank, 2022). There are several motivations for this perspective.

First, a central desideratum in designing feature systems is to allow them to reference all and only the classes of sounds that pattern together cross-linguistically: namely, those that share some subset of phonetic properties encoded by the feature system. However, linguists have discovered a substantial number of phonological classes across languages that cannot be referenced under standard feature systems (Mielke, 2008). An example of one such class is the segments that participate in a nasalization process in Evenki (Tungusic; Nedjalkov, 1997; Mielke, 2008): the sounds /v s g/ become nasalized following a nasal consonant, but similar sounds such as /b d x/ do not. It is not possible to provide a set of feature/value pairs that picks out the class /v s g/ to the exclusion of all other sounds in the language, which predicts that it should not pattern cohesively. In similar cases, researchers have proposed modifications to existing feature systems to account for unexpected classes (though perhaps not modifications so extreme as to capture /v s g/; e.g., Rice and Avery, 1989; McCarthy, 1991; Paradis and LaCharité, 2001).

Emergent feature theory instead proposes that features may be learned in part from the distributional patterning of sounds, which means a shared representation could be learned for irregular classes like /v s g/ if the language data supported it. This also turns the focus away from enumerating all of the features motivated by natural language phonology, focusing instead on how features might be learned from the phonetic and distributional properties of sounds.

A second, related, motivation for emergent features is the variable patterning of the same segment across different languages. For example, Mielke (2008) notes that some languages treat /l/ as [+continuant], and others treat it as [−continuant]. Both are sensible from the perspective of phonetic substance, since /l/ is [-continuant] mid-sagittally but [+continuant] off mid-line. Rather than trying to determine the "correct" value of [continuant] for /l/, or perhaps to split [continuant] into a pair of features corresponding to on and off the mid-line, emergent feature theory suggests that the featural representation of /l/ can vary depending on whether it patterns with [+continuant] or [−continuant] sounds in a language.

Several computational models have been proposed to test the plausibility of distributional learning of phonological classes/features (e.g., Goldsmith and Xanthos, 2009; Mayer, 2020; Nelson, 2022). These papers have tested phonological class learning under the extreme assumption that the learner has no access to the substantive phonetic properties of segments, but only their statistical patterning. Representations learned from distribution alone have been shown to capture non-trivial phonetic distinctions as well as distribution-specific information (Goldsmith and Xanthos, 2009; Mayer, 2020) and to perform comparably to phonetic features in downstream tasks (Nelson, 2022).

The segmental representations in such models are learned using similar techniques to distributional word embeddings (Mikolov et al., 2013; Levy and Goldberg, 2014), which produce real-valued vector representations. In phonological theory, features serve as an extensional description of phonological classes, and most models of phonotactics assume discrete features accordingly. A common feature of the models above is that they use clustering techniques to convert these continuous representations into discrete classes. These classes can then be converted into discrete featural representations (Mayer and Daland, 2020).

Although the process of converting continuous representations to discrete ones aligns with the standard theoretical treatment, it discards information and introduces additional degrees of freedom into

the learning process, in the sense that choices must be made about how clustering is done and how features are derived from classes. Several neural models of phonotactics have used continuous representations directly (Mirea and Bicknell, 2019; Mayer and Nelson, 2020). These recurrent neural network models perform well but are difficult to interpret in a theoretically-satisfying way because they involve many nonlinear transformations of the input features.

1.3 Overview of this paper

This paper presents a computational model[1] that bridges the gap between distributional learning techniques and phonotactic models by incorporating the induction of continuous distributional representations into the overall framework of phonotactic learning. More specifically, we will show that (a) the proposed model is flexible enough to make use of a range of different featural representations, including the continuous features typically produced by distributional learning techniques; (b) the model performs comparably to other models in the field; and (c) the continuous distributional representations result in better generalization to new data than their discretized counterparts, and outperform phonetic features in some respects.

Sections 2 and 3 describe the proposed model and three types of featural representation that will be used to test the model. Section 4 presents a simple toy example to illustrate the performance of the model, and Sections 5 and 6 compare the performance of the model on English onsets against several other models of phonotactic learning. Section 7 offers a brief discussion.

2 Model description

Our goal is to develop a model for the probability of a form in terms of the conditional probability of a symbol x given its preceding context c, in a way that leverages potentially real-valued featural representations of x and c, such as those resulting from distributional analysis, without needing to reduce these continuous representations into hard categories or clusters. To these ends, we adopt **log-bilinear** probability models, a generalization of the widely used log-linear model. Below, we first describe log-linear models and their relation to existing models of phonotactics, then their generalization to log-bilinear models.

2.1 Log-linear models

In a log-linear model, a form is assigned a probability as a function of weighted features.[2] One example is the Maximum Entropy phonotactic model proposed by Hayes and Wilson (2008), in which a wordform x is described in terms of a constraint violation profile: a vector $\phi(x)$ whose values are the number of times the wordform violates each constraint. The probability of x under the model is then

$$p(x) \propto \exp\Big\{\mathbf{w}^\top \phi(x)\Big\}, \tag{1}$$

where the weight vector $\mathbf{w}$ represents the weight of each constraint. The vector $\mathbf{w}$ is found by optimization to maximize the likelihood of a given dataset of forms.

Such models are called *log-linear* because the function in Eq. 1 is linear after taking a logarithm. In the context of phonotactics, the linear component of this model is a Harmonic Grammar model (Smolensky and Legendre, 2006; Pater, 2009) that uses numerical constraint weights and assigns each word a numerical score based on its violation profile. Log-linear models are one way of using these scores to compute a probability distribution over words (cf. Boersma and Pater, 2016).

Log-linear models are ubiquitous not only in computational learning models but also in natural language processing (e.g., Berger et al., 1996; Della Pietra et al., 1997). Before the modern renaissance of neural networks, the dominant paradigm for any supervised classification task in NLP (for example, the task of reading in a movie review and then outputting the probability that the review is positive) was to use a hand-crafted featural representation $\phi(x)$ of the text input x and to learn optimized weights $\mathbf{w}$ to maximize the likelihood of labels in training data (Jurafsky and Martin, 2023).

2.2 The current proposal: log-bilinear model

The log-bilinear model extends the log-linear model to make the weights conditional on the features of the context. Instead of finding an optimal weight vector, in a log-bilinear model one finds an optimal weight *matrix* that relates the representations of the context to the representations of the outcome. Such models were initially developed in a

[1] The code and data used in this paper can be found at https://github.com/hutengdai/vector_bilinear.

[2] Features in this context refer to properties of the form in general, not necessarily phonological features.

language modeling context to predict words given previous words (Mnih and Hinton, 2007, 2008; Mikolov et al., 2013; Futrell, 2022).

We apply a log-bilinear model in the setting of calculating the conditional probability of an individual segment x conditional on a context c, given vector representations of the segment $\phi(x) \in \mathbb{R}^K$ and of the context $\psi(x) \in \mathbb{R}^L$. The model is defined as

$$p(x \mid c) \propto \exp\Big\{\psi(c)^\top \mathbf{A}\phi(x)\Big\}, \qquad (2)$$

where $\mathbf{A} \in \mathbb{R}^{K\times L}$ is an **interaction matrix** that defines how the features of the context $\psi(c)$ relate to the features of the result $\phi(x)$. The entry A_{kl} in the interaction matrix is an association weight for the kth feature of the context and the lth feature of the next segment; a high value of A_{kl} means (all else being equal) that a segment with a high value of the lth feature is likely to follow in a context with a high value of the kth feature.

The interaction matrix $\mathbf{A}$ is found to maximize the likelihood of a training dataset consisting of N context–outcome pairs $\{\langle c_n, x_n\rangle\}_{n=1}^N$:

$$\mathbf{A} = \arg\max_{\mathbf{A}} \sum_{n=1}^{N} \log p(x_n \mid c_n). \qquad (3)$$

The implemented learning algorithm discovers the interaction matrix using the Adam optimization algorithm (Kingma and Ba, 2015), starting from a randomly-initialized $\mathbf{A}$ whose entries are all drawn from a standard Normal distribution.

We model the likelihood of a wordform in terms of features of segmental bigrams. That is, the weights learned by the model correspond to the strength of bigram constraints on the features of two adjacent segments. The probability for a form $\sigma_1, \ldots, \sigma_T$ is then

$$p(\sigma_1, \ldots, \sigma_T) = \prod_{t=1}^{T} p(\sigma_t \mid \sigma_{t-1}), \qquad (4)$$

where $p(\cdot \mid \cdot)$ is a log-bilinear model with the same featurization $\phi(\cdot)$ for the current segment σ_t and the context σ_{t-1}. This restriction to featural bigram constraints is an implementation detail; the log-bilinear model works with any vector representation of context and target. In particular, context and target representations do not need to be the same size; the context representation can include information about multiple segments by increasing the dimension of $\psi(c)$ and $\mathbf{A}$ accordingly.

3 Featurizations

We will illustrate the performance of the log-bilinear model described above using three types of featurizations that have been used in the literature on phonotactic learning: **discrete phonetic** features, **continuous distributional** features, and **discrete distributional** features. The purpose of these comparisons is to (a) demonstrate the flexibility of the model in terms of representational choices; and (b) show that the continuous distributional representations contain useful, fine-grained information that is lost when these representations are discretized.

3.1 Discrete phonetic features

An obvious choice for the featurization of a segment σ is the discrete phonetic features that are commonly used in phonological theory. We adopt the featurization system from Hayes (2009).

For models where featural representations are treated as numerical vectors, such as the log-bilinear model, we adopt a **binary featurization** that identifies each dimension of $\phi(\sigma)$ with a phonological feature *and its possible values*. So for example, there would be a separate dimension for the feature-value pairs $[+\text{continuous}]$ and $[-\text{continuous}]$ with value 1 if that feature-value pair applies to the segment σ and 0 otherwise. For example, the segment [k] would receive the vector representation

$$\phi(\text{k}) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \begin{matrix} +\text{dorsal} \\ -\text{dorsal} \\ +\text{continuous} \\ -\text{continuous} \\ +\text{consonantal} \\ -\text{consonantal} \\ \vdots \end{matrix}. \qquad (5)$$

This leads to a more expressive featurization than encoding negative values as -1. This would force the effect of a negative feature value to be the inverse of the effect of a positive feature value, whereas the binary featurization allows positive and negative values to have independent effects.

3.2 Continuous distributional representations

We induce continuous representations based on their statistical distributions in the training data by calculating probabilities of segments in different contexts and then converting these into

Pointwise Mutual Information (PMI; Church and Hanks, 1990). PMI is an information-theoretic measurement that compares the joint probability of two events against the product of their individual probabilities. PMI and the related Positive PMI have been used in previous models of distributional phonotactic learning (Silfverberg et al., 2018; Mayer, 2020; Nelson, 2022).

PMI is defined as follows:

$$\text{PMI}(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}. \quad (6)$$

If $p(x)$ and $p(y)$ are independent this value will be close to zero, while if they occur together more/less frequently than chance, it will be positive/negative. Here we define $p(x, y)$ to be the joint probability of segment x followed by segment y. We compute the probabilities using a bigram language model with Kneser-Ney smoothing (Chen and Goodman, 1999), implemented using the `lm` module from the `nltk` Python library (Bird et al., 2009). This model produces conditional probabilities of the form $p(y|x)$, which we convert to joint probabilities $p(x, y)$.

The dimensions of these representations are the PMI values of the segment in each context in the training data. Following Mayer (2020), we consider both preceding and following context by running a pair of language models: one that runs forward to calculate PMI values based on preceding context, and one that runs backwards to calculate PMI values based on following context. These two vectors are concatenated to produce the full representation.

3.3 Discrete distributional features

We also include discrete distributional featurizations derived from the continuous representations in the previous section. This discretization step allows the distributional representations to be used in models that assume discrete features.

Converting continuous features to discrete ones involves two steps: a *clustering step* where classes of segments are identified based on similarities in their continuous representations, and a *feature assignment step* where a feature system is derived from these classes.

We include two clustering strategies: the recursive clustering algorithm described in Mayer (2020) and the SC COV algorithm from Nelson (2022). Both of these involve using the continuous embeddings to compute graph structures that reflect distributional similarity between segments, and then applying graph partitioning techniques to derive classes of segments. For reasons of space we refer the reader to the respective papers.

We follow Nelson (2022) in using the *inferential complementary* algorithm from Mayer and Daland (2020) for feature assignment. Mayer and Daland (2020) presents a suite of algorithms that derive a feature system from a set of input classes based on subset/superset relationships between them, differing in what values are permitted and whether complement classes of the input classes are inferred. The inferential complementary algorithm adds complement classes of the input classes with respect to their parents and assigns both $+$ and $-$ feature values.[3]

4 A toy example of the log-bilinear model

We present a simple toy example below to illustrate the performance of the log-bilinear model using the continuous distributional features described in Section 3.2. We define a language over the alphabet {C, V, #}, where # is a word boundary. The language has a restriction on adjacent CC sequences, and the training data is {VCV, CVC, CVV, VVC, VVV} (word boundaries are omitted for clarity). The continuous distributional featurization of each segment calculated from the training data is shown in Table 1. Sequences that are unattested in the training data, such as ## or CC, have large negative scores, while more commonly observed contexts have positive scores.

	#_	C_	V_
#	-2.492	0.504	0.232
C	0.517	-3.256	0.251
V	0.111	0.118	-0.278

Table 1: Continuous distributional representations of the segments in the toy language. Each row is a representation of a segment, and the columns are the PMI values of that segment in the context indicated by the column label. For simplicity's sake we only present preceding contexts here, but the full model also includes dimensions corresponding to following context.

[3] Nelson (2022) in fact uses a slightly simplified version of this algorithm: the original algorithm recursively adds complement classes until there are no more to add, while the algorithm in Nelson (2022) adds complement classes once and then terminates. This potentially reduces the expressivity of the feature system, but the two approaches are similar enough that we treat them as a single feature assignment strategy.

Table 2 shows the scores assigned by the log-bilinear model to a set of nonce words after it was fitted to the training data using the representations in Table 1. The model successfully assigns a lower probability to words containing a CC sequence.

Word	Score
C V C V	5.397
V C V V	5.980
V V V V	6.393
C C C V	8.825
V C C C	8.933
C C C C	10.272

Table 2: Scores assigned by the trained model to nonce forms. The scores here are negative log probabilities.

5 Model comparison

We evaluate the performance of the log-bilinear model against several existing models of phonotactics. These models take as input a set of training data and, in most cases, a set of featural representations for the segments in the training data. Fitted models assign scores to word forms that reflect their probabilities.

The purpose of this comparison is to demonstrate that the log-bilinear model performs favorably against existing phonotactic models.

5.1 Hayes and Wilson learner

The Hayes and Wilson learner (Hayes and Wilson, 2008) is a Maximum Entropy model of phonotactics. We refer the reader back to Section 2.1 for a description of how word probabilities are computed based on input constraint violation profiles and a set of learned weights.

In addition to fitting weights, the Hayes and Wilson learner also simultaneously learns the constraints themselves from the data, up to an upper bound specified by the user. Constraints are implemented as featural n-gram constraints (e.g., *[−voi, −son][+voi, −son]). Constraints are discovered by comparing observed vs. expected counts in the training data and selecting constraints that penalize structures with unexpectedly low counts. There is a bias towards constraints that include fewer features, but more complex interactions are learned when the data support them.

The scores assigned by this model are *harmony values*, which are unnormalized log probabilities (the linear component of the log-linear model).

5.2 MaxEntGrams

MaxEntGrams[4] is a variant of the Hayes and Wilson learner that offers time and space improvements over the original algorithm by training on an n-gram model of the training data rather than the data itself. For a more detailed comparison of the two models, see Nelson (2022). This model also produces unnormalized log probabilities.

5.3 Smoothed bigram model

This model is included as a baseline. It defines the probability of a word as in Eq. 4, but with conditional probabilities estimated from counts with additive smoothing:

$$p(\sigma_t|\sigma_{t-1}) = \frac{C(\sigma_{t-1}, \sigma_t) + 1}{C(\sigma_{t-1}) + d}, \qquad (7)$$

where $C(\sigma_{t-1}, \sigma_t)$ is the count of the sequence $\sigma_{t-1}\sigma_t$ in the training data, $C(\sigma_{t-1})$ the count of σ_{t-1}, and d the number of distinct segments. This score is reported as a log probability.

This model operates on segmental representations, and thus cannot generalize along featural dimensions. Additive smoothing mitigates this somewhat by assigning every segmental bigram an initial pseudo-count of 1. This ensures that forms containing bigrams that are not in the training data are assigned low, rather than zero, probabilities.

5.4 Summary of models

We do not consider every possible permutation of the models and featurizations above, but present the set shown in Table 3. In particular, we report only a single combination of the models presented in Nelson (2022). In addition to comparing the models themselves, we also focus our analysis on the dimensions of *continuous vs. discrete features* and *phonetic vs. distributional features*.

6 Model comparison on English onset sequences

We compare the performance of the log-bilinear model against the models above on the problem of learning restrictions on onset clusters in English. This problem has been extensively studied in the context of the **Sonority Sequencing Principle** (SSP): the cross-linguistic preference for syllable onsets that monotonically increase in sonority and codas that monotonically decrease in sonority

[4] https://github.com/MaxAndrewNelson/PhoneGraphs

Model	Featurization
Smoothed bigram	N/A
Hayes & Wilson	Discrete phonetic
Hayes & Wilson	Discrete distributional (Mayer)
Bilinear	Continuous distributional (PMI)
Bilinear	Discrete phonetic
Bilinear	Discrete distributional (Mayer)
MaxEntGrams	Discrete distributional (SC COV)

Table 3: Models to be tested

(Selkirk, 1984). Sensitivity to the SSP has been found in many experimental studies, and it has been argued that it constitutes an innate phonological bias (Berent et al., 2008, 2011). Computational studies have shown that phonotactic learning models operating on lexical statistics can learn generalizations about the SSP that align with human behavior, despite having no biases towards SSP-conforming onsets (Daland et al., 2011; Mayer and Nelson, 2020; Nelson, 2022). However, models that incorporate both prior bias and statistical learning have been shown to account better for SSP judgments than either does individually, suggesting a role for both bias and experience (Jarosz and Rysling, 2017; Jarosz, 2017/8). We do not employ this dataset here to make any strong claims about the innateness of the SSP, but rather because it has been used to compare the performance of phonotactic models in previous work.

The training data for all models was the English onset corpus from Hayes and Wilson (2008). This consists of all word-initial onsets from the CMU Pronouncing Dictionary (Weide et al., 1998, `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`): thus each word type in the dictionary contributes a single token to the corpus. Hayes and Wilson sanitize the corpus by removing "exotic" onsets such as [zw], [sf], and [pw] that are unlikely to be encountered by language learners, and by assuming that [j] off-glides are part of the nucleus. We used this dataset to construct the distributional embeddings and to fit the parameters of each model. Following Nelson (2022), the distributional embeddings were calculated over the set of unique onsets (or onset types).

We did a hyperparameter search using cross-validation to determine the learning rate and batch size used to train the log-bilinear model. We considered the values $[32, 64, 128, 256, 512, 1024, 2048, 4096]$ for batch size and $[0.1, 0.01, 0.001, 0.0001]$ for the learning rate. A batch size of 64 and learning rate of 0.001 led to the optimal fit.

We restricted the H&W learner to bigram constraints, allowed it to learn a maximum of 300 constraints, and used the default maximum Observed/Expected threshold of 0.3.

The models were tested on the experimental data from Daland et al. (2011). These data consist of Likert ratings given by 48 participants to a set of 96 nonce words beginning with 48 different onsets. Daland et al. (2011) group the onsets into three different classes: *attested* onsets, which are common in English, *marginal* onsets, which are attested but uncommon, and *unattested* onsets. Following Nelson (2022), we train and test on the onsets in isolation (i.e., the data consist of forms like "sm", "pl", etc.). Each onset is represented by two data points corresponding to two tails the onset was attached to in the Daland et al. study. The onsets are shown in Table 4.

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

Table 4: Onsets from Daland et al. (2011).

The trained models assigned scores to the test data according to their onsets. We evaluated model performance by looking at the correlation of scores assigned by each model to the Likert ratings provided by human participants. Following Daland et al. (2011), we look at correlations within the attested/marginal/unattested onset groups, as well as overall correlation. We report both Pearson's r, which captures relative differences in well-formedness but is sensitive to non-linearity between model scores and human judgments, and Kendall's τ, which is not sensitive to non-linearity but only considers the rank ordering of points (see Albright, 2009).

The results are shown in Table 5. The two most successful models are the Hayes & Wilson learner with discrete phonetic features, and the log-bilinear model with continuous distributional features: these have the two highest overall τ correla-

Model	Featurization	Overall		Attested		Marginal		Unattested	
		r	τ	r	τ	r	τ	r	τ
Smoothed bigram	segments	**0.877**	0.669	0.509	0.244	0.274	-0.004	0.470	0.280
MaxEntGrams	discrete dist.	0.753	0.610	0.424	0.282	0.212	0.171	**0.583**	0.417
H&W	discrete phon.	0.740	0.674	0.533	0.261	**0.422**	**0.301**	0.459	0.374
	discrete dist.	0.818	0.634	0.540	0.244	-0.012	-0.049	0.547	0.421
Bilinear	discrete phon.	0.785	0.646	0.446	0.215	0.367	0.247	0.525	0.377
	discrete dist.	0.757	0.572	0.520	0.296	0.021	0.067	0.523	0.309
	continuous dist.	0.699	**0.694**	**0.611**	**0.332**	0.247	0.201	0.562	**0.465**

Table 5: Model comparison using Pearson's r and Kendall's τ to correlate model scores with acceptability ratings for English onsets. The correlation value for the top performing model in each category is bolded.

tions and achieve the highest τ correlations in each of the four categories. Fig. 1 shows the relationship between model scores and human Likert ratings.

The high performance of the bilinear model with continuous distributional features when compared against the same model with discretized distributional features shows that the continuous features contain phonotactically relevant information which is lost under discretization.

It is also interesting to note that the distributional models achieve the highest correlations for all but the marginal forms, which are best captured by models with phonetic features. This may suggest that the relative importance of distributional vs. phonetic information varies in different contexts, but more research will be needed to see if this observation is borne out more generally.[5]

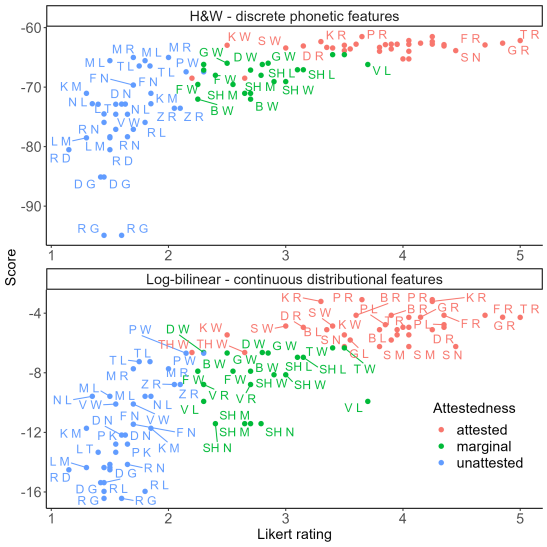

Figure 1: Comparison of the predictions of the two most successful models against human Likert ratings.

7 Conclusion

This paper has presented a log-bilinear model of phonotactics that can incorporate continuous representations of phonological information, bypassing the discretization steps used in previous work. The results of a modeling study showed that this model achieves competitive performance in predicting experimental judgments of English onsets. This model opens up the space of possibilities for phonotactic modeling by removing requirements for discrete representations, allowing greater compatibility with standard distributional learning techniques.

The log-bilinear model is also compatible with continuous representations proposed in other contexts, such as on the basis of phonetic measurements (Mielke, 2012). The model could be used to implement a model of phonotactics that operates directly on these representations, providing insight into the role of fine-grained phonetic detail in phonotactic judgments. More generally, different feature systems may be compared within the log-bilinear framework, and the log-bilinear model itself can be used to generate optimized distributional vector representations of segments: this is the method used to create word2vec vectors when applied to text data (Mikolov et al., 2013; Goldberg and Levy, 2014).

Finally, the log-bilinear model can be straight-

[5]The high performance of the smoothed bigram model on the overall Pearson's correlation is likely due to a strong numerical match with the acceptability ratings of the attested forms, as noted by Daland et al. (2011): performance on unattested and marginal categories, and using Kendall's τ, is substantially worse.

forwardly applied to larger contexts than bigram windows, including autosegmental or tier-based contexts (Goldsmith, 1976; Heinz et al., 2011), by appropriately defining the context representation. The flexibility, relative simplicity, and performance of this model make it a promising framework for studying phonotactic learning and representations.

References

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

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.

Diana Archangeli and Douglas Pulleyblank. 2022. *Emergent phonology*. Language Science Press, Berlin.

Iris Berent, Katherine Harder, and Tracy Lennertz. 2011. Phonological universals in early childhood: Evidence from sonority restrictions. *Language Acquisition*, 18(4):281–293.

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.

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

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural language processing with Python*. O'Really Media Inc.

Paul Boersma and Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In John J. McCarthy and Joe Pater, editors, *Harmonic Grammar and Harmonic Serialism*. Equinox, Sheffield.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Noam Chomsky and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of Linguistics*, 1(2):97–138.

Kenneth W. Church and Patrick Hanks. 1990. Word association, norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

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

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

Stephen A. Della Pietra, Vincent J. Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions: Pattern Analysis and Machine Intelligence*, 19:380–393.

Richard Futrell. 2022. Estimating word co-occurrence probabilities from pretrained static embeddings using a log-bilinear model. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 54–60, Dublin, Ireland. Association for Computational Linguistics.

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 natural classes. *Phonology*, 36(1):37–60.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

John Goldsmith. 1976. *Autosegmental phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

John Goldsmith and Aris Xanthos. 2009. Learning phonological categories. *Language*, 85(1):4–38.

Maria Gouskova and Gillian Gallagher. 2020. Inducing nonlocal constraints from baseline phonotactics. *Natural Language and Linguistic Theory*, 38(1):77–116.

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.

Bruce Hayes, Kie Zuraw, Peter Siptar, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language*, 85:822–863.

Jeffrey Heinz, Chetan Rawal, and Herbert G. Tanner. 2011. Tier-based strictly local constraints in phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 58–64.

Gaja Jarosz. 2017/8. Defying the stimulus: acquisition of complex onsets in Polish. *Phonology*, 34(2):269–298.

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.

Daniel Jurafsky and James H. Martin. 2023. Speech and language processing (3rd edition). https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Connor Mayer. 2020. An algorithm for learning phonological classes from distributional similarity. *Phonology*, 37(1):91–131.

Connor Mayer and Robert Daland. 2020. A method for projecting features from observed sets of phonological classes. *Linguistic Inquiry*, 51(4):725–763.

Connor Mayer and Max Nelson. 2020. Phonotactic learning with neural language models. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 291–301, New York, New York. Association for Computational Linguistics.

John J. McCarthy. 1991. Semitic gutturals and distinctive feature theory. *Perspectives on Arabic linguistics*, 3:63–91.

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

Jeff Mielke. 2012. A phonetically based metric of sound similarity. *Lingua*, 122(2):145–163.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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

Andriy Mnih and Geoffrey E Hinton. 2007. Three new graphical models for statistical language modelling. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, pages 641–648.

Andriy Mnih and Geoffrey E Hinton. 2008. A scalable hierarchical distributed language model. *Advances in Neural Information Processing Systems*, 21:1081–1088.

Igor Nedjalkov. 1997. *Evenki*. Routledge, London.

Max Nelson. 2022. *Phonotactic learning with distributional representations*. Ph.D. thesis, University of Massachusetts, Amherst.

Carole Paradis and Darlene LaCharité. 2001. Guttural deletion in loanwords. *Phonology*, 18(2):255–300.

Joe Pater. 2009. Weighted constraints in generative linguistics. *Cognitive Science*, 33:999–1035.

Karen Rice and Peter Avery. 1989. On the interaction between sonorancy and voicing. *Toronto Working Papers in Linguistics*, 10.

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*.

Paul Smolensky and Geraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality theoretic grammar*. MIT Press, Cambridge.

Robert Weide et al. 1998. The Carnegie Mellon pronouncing dictionary. *Release 0.6, www.cs.cmu.edu*.

Colin Wilson and Gillian Gallagher. 2018. Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic Inquiry*, 49(3):610–623.