Reconciling categorical and gradient models of phonotactics

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

Abstract

Should phonotactic knowledge be modeled as categorical or gradient? In this paper, I present new data from a Turkish acceptability judgment study that addresses some limitations of previous work on this question. This study shows that gradient models account for the variability in acceptability ratings better than categorical ones. However, I suggest that the distinction between gradient and categorical models is somewhat superficial when we think of models in a mathematically general way. I propose on this basis that both categorical and gradient models have a role to play in linguistic research.

1 Is phonotactics gradient or categorical?

Phonotactics is the restrictions that languages place on how sounds can be sequenced into words. Different languages impose different phonotactic restrictions. For example, although English and Spanish both contain the sounds {k, p, s, i}, a word like /skip/ 'skeep' is only possible in English. Spanish has more restrictive phonotactics, prohibiting /s/-initial complex onsets. For similar reasons, a word like /fstʃɔŋs/ is a perfectly fine Polish word (*wstrząs* 'shock'), but would not be a suitable English word because of English's more restrictive onset phonotactics. It is generally accepted that phonotactic knowledge is learned by generalizing across forms in the lexicon (e.g. Chomsky and Halle, 1968; Bailey and Hahn, 2001; Edwards et al., 2004).

One common method of probing phonotactic knowledge is phonotactic acceptability judgments, where participants are asked to rate the acceptability of novel words as possible words in their language. A longstanding empirical observation is that phonotactic acceptability judgments are *gradient*. That is, participants do not simply treat words as acceptable or not, but rather ascribe varying degrees of acceptability to them. A classic example from Chomsky and Halle (1968) is the three nonce words /blɪk/, /bnɪk/, and /bnzk/. Despite all three being unattested in English, English speakers (or at least Chomsky and Halle) rank them in terms of acceptability such that /bnzk/ $\ll$ /bnɪk/ $\ll$ /blɪk/. That is, speakers judge /bnɪk/ to be a more acceptable word than /bnzk/, but a less acceptable word than /blɪk/. Similar results have been found in a wide range of studies (e.g. Coleman and Pierrehumbert, 1997; Scholes, 1966; Hayes, 2000; Bailey and Hahn, 2001; Hayes and Wilson, 2008; Albright, 2009; Daland et al., 2011, a.o.).

Two question that naturally arise from these results are where this gradience comes from and how we should represent it in our models of language. There have been two broad theoretical approaches, which we will cover in the following sections (see Schütze, 1996, for a discussion of these perspectives in linguistics more broadly).

1.1 Gradient models of phonotactics

The first approach proposes that we see gradience in these studies because the phonotactic grammar is itself gradient, or that a gradient measure of acceptability can be derived from the grammar. Chomsky and Halle (1968) write that "a real solution to the problem of 'admissibility' will not simply define a tripartite categorization of occurring, accidental gap, and inadmissible, but will define the 'degree of admissibility' of each potential lexical matrix in such a way as to distinguish /blɪk/ from /bnɪk/ and /bnɪk/ from /bnzk/, and to make numerous other distinctions of this sort" (pp. 416–417). They operationalize this 'degree of admissibility' as a quantity derived from the phonological grammar and the lexicon: the minimum number of featural changes required to convert a word into an existing word in the language. Chomsky and Halle also note that this gradience exists within the lexicon itself (p. 418). In English, for example, there are semi-admissible words like /sfɪŋks/ 'Sphinx' that constitute exceptions to otherwise strong phonotactic restrictions on onset formation.

Proceedings of the Society for Computation in Linguistics (SCiL) 2025, pages 144-154.
Eugene, Oregon, July 18-20, 2025

Chomsky and Halle do not do away with the concept of grammaticality: there are still forms that can be produced by the grammar and forms that cannot. Rather, they suggest that a gradient acceptability score can be derived from the grammar by some additional mechanism. Subsequent proposals have gone further, claiming that the grammar itself generates both categorical and gradient outcomes: whether we get one or the other depends primarily on the amount of variability in the learning data. It's beyond the scope of this paper to cover these approaches in detail, but many have been expressed within the context of Optimality Theory (Prince and Smolensky, 1993/2004) and typically either vary constraint rankings in order to generate gradient outcomes (e.g. Hayes, 2000) or derive probabilities from weighted constraints (e.g. Hayes and Wilson, 2008; Dai et al., 2023). Gradient models of phonotactics have also been proposed in the context of formal language theory (Mayer, 2021). Under these approaches, gradience emerges from an interaction between the grammar and the learning data, not a bespoke mechanism.

This perspective is supported outside the world of generative linguistics, where phonotactic knowledge is typically treated as gradient, and is often represented by simple probabilistic n-gram models (Markov, 1913; Shannon, 1948). Gradient knowledge of phonotactics has been claimed to play an important role in areas such as speech perception (e.g. Norris and McQueen, 2008; Dupoux et al., 2011; Chodroff and Wilson, 2014; Steffman and Sundara, 2023), speech production (e.g. Edwards et al., 2004), word segmentation and learning (e.g. Mattys et al., 1999; McQueen, 1998; Mersad and Nazzi, 2011; Vitevitch and Luce, 1999; Storkel, 2001), and speech errors (e.g. Goldrick and Larson, 2008; Taylor and Houghton, 2005; Warker, 2013; Warker and Dell, 2006, 2015), among others.[1]

1.2 Categorical models of phonotactics

The second theoretical approach to gradience proposes that the phonotactic grammar is fundamentally categorical (that is, it really does judge words to be acceptable or not) and that gradience in acceptability judgments is solely the result of extra-grammatical factors such as task effects or misperception (e.g. Gorman, 2013; Durvasula, 2020; Kostyszyn and Heinz, 2022; Dai, 2025). There are two main sources of evidence for this view.

The first is that extra-grammatical performance factors have indeed been shown to influence phonotactic judgments. A convincing demonstration of this comes from Kahng and Durvasula (2023), who show that some variability in nonce word judgments by Korean speakers is the result of misperception of certain consonant clusters.

The second source of evidence is several studies suggesting that categorical models do as well as or better than gradient models in predicting acceptability judgments. As Gorman (2013) puts it, "simple baselines better account for gradient well-formedness judgements than current computational models of phonotactic knowledge, suggesting that the gradience observed in these tasks [does] not derive from known grammatical mechanisms" (p. 17). Specifically, categorical models have been claimed to better predict English onset acceptability (Gorman, 2013; Durvasula, 2020; Dai, 2025), Polish onset acceptability (Kostyszyn and Heinz, 2022; Dai, 2025), Turkish vowel harmony (Gorman, 2013; Dai, 2025) and English medial consonant cluster distributions (Gorman, 2013).

We will focus on the second type of evidence here. With regards to the first, note that proponents of gradient models do not suggest that extra-grammatical factors have no role at all in the gradience exhibited in acceptability judgment tasks. Rather, the claim is that a substantial part of the gradience can be predicted by grammatical factors. Hayes (2000) puts it as follows:

> [P]atterns of gradient well-formedness often seem to be driven by the very same principles that govern absolute well-formedness [. . .] I conclude that the proposed attribution of gradient well-formedness judgments to performance mechanisms would be uninsightful. Whatever "performance" mechanisms we adopted would look startlingly like the grammatical mechanisms that account for non-gradient judgments (p. 90).

In other words, gradience in acceptability studies is often predictable from "soft" versions of the same constraints that govern more categorical patterns like phonological alternations.

[1] We do not consider neighborhood density here, another important property that influences wordlikeness judgments. For discussion of the relationship between neighborhood density and phonotactic probability, see e.g. Bailey and Hahn (2001); Steffman and Sundara (2024).

1.3 Limitations of past work

There are three important limitations to previous work comparing categorical and gradient models of phonotactics. First, these papers have used a relatively small number of data sets, almost all focusing on consonant clusters. This makes it difficult to evaluate how generally these results hold across different types of phonotactic dependencies.

The second limitation is that the authors of these papers do not all subscribe to the same definition of categorical. In some cases the grammar truly is categorical, assigning words either grammatical or ungrammatical status (Gorman, 2013; Kostyszyn and Heinz, 2022; Dai, 2025). In other cases, similar to Chomsky and Halle (1968), some secondary gradient measure of admissibility is derived from a categorical grammar (Durvasula, 2020; Kostyszyn and Heinz, 2022). We will treat these two definitions of categorical as separate models below.

The third limitation is that the gradient model typically used is the UCLA Phonotactic Learner (Hayes and Wilson, 2008), an influential phonotactic learning model implemented in the maximum entropy Optimality Theory framework (Goldwater and Johnson, 2003; Mayer et al., 2024). Although it does implement a gradient model of phonotactics, it has the additional task of inducing the constraints themselves from the data. The categorical models in these papers are typically provided with predefined constraints (though cf. Dai, 2025). It is unclear whether the poor performance of the UCLA learner is due to the fact that it is gradient or to some aspect of the constraint induction process. The UCLA learner is also sensitive to how it is parameterized, and it is not typical for these studies to compare performance under a range of hyperparameters.

1.4 The remainder of the paper

While this paper will by no means resolve this debate, I will try to achieve two more modest goals. First, I will present new data from a phonotactic acceptability judgment study of Turkish that addresses some of the limitations expressed above. This study will show that gradient models are better able to predict participant judgments. Second, I will try to convince you that the distinction between categorical and gradient grammars is in fact a somewhat superficial one when we consider the matter from a mathematical perspective, and that both conceptualizations of the grammar have a role to play in linguistic research and theory-building.

2 Defining our grammars

We will consider three classes of models in the rest of the paper. *Boolean* models, *cost* models, and *probability* models. Abstracting away from the internal details for a moment, we can think of each of these models as defining a `score` function that assigns some value to a string:

$$\text{score} : \Sigma^* \rightarrow \mathcal{T}$$

where Σ is a set of symbols, Σ^* is the set of all possible strings generated from this set, and $\mathcal{T}$ is some set of values. The three models differ in what type of value the `score` function assigns.

2.1 Boolean models

We will use boolean models to correspond to the theoretical position that the phonotactic grammar is categorical, with gradience stemming from non-grammatical factors (Gorman, 2013; Kostyszyn and Heinz, 2022; Dai, 2025). The score function for these models assigns boolean values to strings:

$$\text{score} : \Sigma^* \rightarrow \{0, 1\}$$

Such models cannot represent a situation where the acceptability of /bnzk/ $\ll$ /bnık/ $\ll$ /blık/. If we take /bnzk/ to be ungrammatical and /blık/ to be grammatical, the model must place the intermediate form /bnık/ into one of these two categories.

2.2 Cost models

Cost models will correspond to the theoretical position that a gradient measure of acceptability is derived from a categorical grammar. There are many ways such a proposal could be implemented, but we will follow Durvasula (2020) and Kostyszyn and Heinz (2022), who derive such a gradient measure by counting the number of (categorical) constraints that a form violates. The `score` function for cost models assigns non-negative integer values to strings, with larger integers corresponding to lower phonotactic acceptability:

$$\text{score} : \Sigma^* \rightarrow \{0, 1, 2, \dots\}$$

In this model, acceptability is bounded on one side by 0, which corresponds to a "perfectly acceptable" form that violates no constraints. The other end of the scale is unbounded, since a form can violate arbitrarily many constraints. This means that, unlike the other two model types, we expect acceptability to *decrease* as the score increases. Such models can represent the case where the acceptability of /bnzk/ $\ll$ /bnɪk/ $\ll$ /blɪk/ by assigning the forms successively decreasing integer values.

2.3 Probability models

Probability models will correspond to the theoretical claim that gradience in acceptability corresponds directly to gradience in the grammar. Gradient grammars do not necessarily have to generate probabilities, but we will assume that is the case here. The score function for probability models is:

$$\text{score} : \Sigma^* \rightarrow [0, 1]$$

Such models can also represent the case where /bnzk/ $\ll$ /bnɪk/ $\ll$ /blɪk/ by assigning the forms successively increasing probabilities.

3 Turkish study

We will compare these three classes of models against new data from a large, online acceptability judgment study of Turkish nonce words.[2] This study expands on a previous acceptability judgment study on Turkish (Zimmer, 1969) by including a much larger number of stimuli and participants and using a slider task rather than a binary forced choice task. We will focus on *backness harmony* and *rounding harmony*, which are common in Turkic languages. Backness harmony requires vowels to agree in backness with the preceding vowel, while rounding harmony requires high vowels to agree in roundness with the preceding vowel (see Table 1). We can implement these restrictions using the following bigram constraints over vowel sequences:

- *[αback] [$-\alpha$back]: a vowel must agree in backness with the preceding vowel.
- *[αround] [$-\alpha$round, +high]: high vowels must agree in roundness with the preceding vowel.

These constraints govern suffix allomorphy: e.g., the plural form of /kedi/ 'cat' is [kedi-ler] 'cat-PL', while the plural of /kuʃ/ 'bird' is [kuʃ-lar] 'bird-PL'. Vowel harmony is is also evident as a strong tendency across the lexicon (though many disharmonic words exist, particularly loanwords) and in acceptability judgment tasks (Zimmer, 1969).

	[−back]		**[+back]**	
	[−round]	**[+round]**	**[−round]**	**[+round]**
[+high]	i	y	ɯ	u
[−high]	e	ø	a	o

Table 1: The vowel system of Turkish

3.1 Methodology

The stimuli consisted of 576 wug words with CVCVC shape. A Python script was used to generate every possible Turkish CVCVC word. Attested words found in the Turkish Electronic Living Lexicon (TELL; Inkelas et al., 2000) were automatically removed. Subsequent manual filtering was done by two native Turkish speakers. The remaining words were scored for unigram and Laplace-smoothed bigram probability using the UCI Phonotactic Calculator (Mayer et al., under revision) based on frequencies from citation forms in TELL. For each unique pair of vowels (8×8 total pairs), nine words were sampled such that they were distributed in a roughly uniform way across the unigram-bigram probability space. As a result, the mean probability of the tokens for each vowel pair was roughly the same (Fig. 1). The 576 tokens were synthesized to speech using Google Cloud. The recordings were vetted by the same two native Turkish speakers for naturalness and clarity.

The experiment was administered using Gorilla (`www.gorilla.sc` Anwyl-Irvine et al., 2020). All materials were presented in Turkish. After providing consent, participants completed a short demographic questionnaire. Participants then completed two screening tasks. The first was an audio check that asked them to identify a word presented to them acoustically. The second was a training run of the main experimental task, where participants were instructed to make a specific selection at the end as an attention check. Failure in either of these tasks led to exclusion from the experiment.

Finally, in the main experimental task, participants were asked to provide acceptability judgments of the stimuli based on their suitability as words in Turkish using a sliding, unnumbered scale. The right side of the scale corresponded to higher acceptability, and high-, mid-, and low-probability words were provided as landmarks (Fig. 2). Stimuli were presented with simultaneous audio and orthographic representation. Slider responses were represented on a numeric scale between 0 and 100, with 100 being the most acceptable.

[2] The data and code for this paper can be found at `https://github.com/connormayer/turkish_phonotactics`

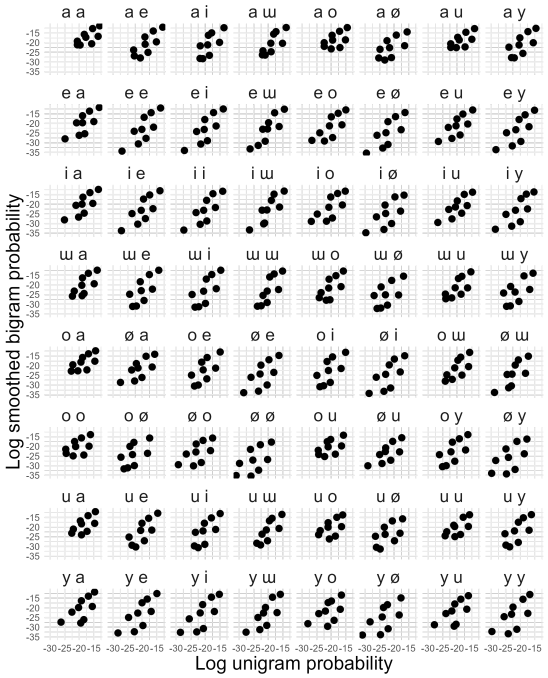

Figure 1: The distribution of unigram and bigram probabilities of the stimuli within each vowel group.

115 native speakers of Turkish were recruited using Prolific (www.prolific.com). 25 participants were excluded because they failed to provide consent or failed one of the two screening tasks. An additional 5 participants were excluded because they indicated in the demographic questionnaire that they had hearing impairment or that Turkish was not their native language. This left a total of 85 participants (38F; mostly age 25–35). Each participant rated 192 tokens after training and attention checks, leading to a total of 16,320 token ratings (about 28 ratings per word). Raw slider responses were normalized to z-scores within participant to control for idiosyncratic differences in mean and spread between participants.

3.2 Results

Fig. 3 shows participant responses broken down by harmonic class. Participants' responses reflect sensitivity to both backness and rounding harmony.

4 Modeling the Turkish data

In this section, we'll compare how well the different models described above predict the acceptability judgment data from the Turkish study. Crucially, each of these models employs the same set of possible constraints, differing only in the values they assign to each. This allows the effect of different value choices to be compared more directly.

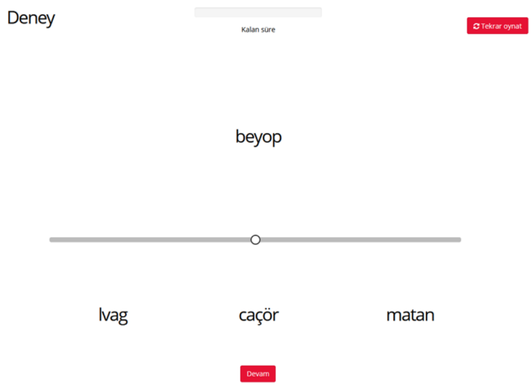

Figure 2: The experimental interface.

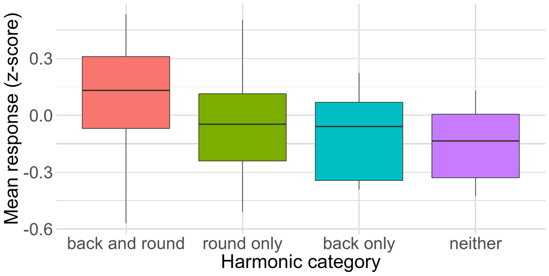

Figure 3: Normalized, mean participant responses broken down by harmonic category. Participants are sensitive to both backness and rounding harmony.

Because our interest is primarily in vowel harmony, we will use tier-based strictly local models with bigram constraints on the vowel tier (a TSL-2 model). It is beyond the scope of this paper to provide a full definition of TSL (see Heinz et al., 2011), but informally it means that we ignore consonants completely and assign scores based only on vowel bigrams. Bigrams can also reference word boundaries (#). This means the models are sensitive not only to which pairs of vowels occur in a word, but also which vowels begin and end the word.

Each model type has a Δ function that assigns a value to a bigram. These bigram values are then aggregated into the value returned by the `score` function discussed above.

4.1 Boolean models

Under a boolean model, the Δ function is:

$$\Delta_b : \Sigma^2 \rightarrow \{0, 1\}$$

where Σ^2 is the set of all possible bigrams, including the word boundary symbol . The boolean values assigned to each bigram in a string are aggregated into a single boolean by conjoining them:

$$\text{score}_b(x_1, \ldots, x_n) = \bigwedge_{i=1}^{n-1} \Delta_b(x_i, x_{i+1})$$

Legal and illegal bigrams receive scores of 1 and 0 respectively. The score for a string is 1 iff it contains only legal bigrams and 0 otherwise.

4.2 Cost models

Under a cost model, the Δ function is:

$$\Delta_c : \Sigma^2 \to \{0, 1, 2 \ldots\}$$

The integers assigned to each bigram are aggregated into a single integer score by summing them.

$$\text{score}_c(x_1, \ldots, x_n) = \sum_{i=1}^{n-1} \Delta_c(x_i, x_{i+1})$$

We will interpret the integer cost assigned to a bigram as the number of bigram constraints it violates. For example, a vowel bigram like /oi/ that violates both backness and rounding harmony might be assigned a cost of 2, while a bigram like /oy/ that violates only backness harmony might be assigned a cost of 1. Although these models could in principle represent varying constraint strengths by assigning different integer costs to each constraint, we will assume following previous work that all constraint violations are equally penalized (Durvasula, 2020; Kostyszyn and Heinz, 2022).

4.3 Probability model

Under a probability model, the Δ function is:

$$\Delta_p : \Sigma^2 \to [0, 1]$$

The probabilities for each bigram are aggregated into a single probability by taking their product:

$$\text{score}_p(x_1, \ldots, x_n) = \prod_{i=1}^{n-1} \Delta_p(x_i, x_{i+1})$$

The individual probabilities assigned to bigrams typically reflect their frequency (though this need not be the case). The probability assigned to a string reflects the probabilities of the bigram sequences it contains.

4.4 An example calculation

Consider again the vowel bigram /oi/. In Turkish, this may be dispreferred because it violates both backness and rounding harmony. Below I show how the score for this sequence can be calculated under each of the three types of models described above (we will discuss where the values assigned to each bigram come from in the following section).

$$\begin{aligned}\text{score}_b(\text{/oi/}) &= \Delta_b(\#\text{o}) \wedge \Delta_b(\text{oi}) \wedge \Delta_b(\text{i}\#)\\ &= 1 \wedge 0 \wedge 1\\ &= 0\end{aligned}$$

$$\begin{aligned}\text{score}_c(\text{/oi/}) &= \Delta_c(\#\text{o}) + \Delta_c(\text{oi}) + \Delta_c(\text{i}\#)\\ &= 0 + 2 + 0\\ &= 2\end{aligned}$$

$$\begin{aligned}\text{score}_p(\text{/oi/}) &= \Delta_p(\#\text{o}) \times \Delta_p(\text{oi}) \times \Delta_p(\text{i}\#)\\ &= 0.08 \times 0.107 \times 0.458\\ &= 0.0004\end{aligned}$$

4.5 Defining Δ

A question that remains is how to actually define Δ for each model: that is, what specific values do we assign to each bigram? We will test several variants that differ in how Δ is defined.

4.6 Δ in the probability model

In the probability model, $\Delta_p(x, y)$ is defined to be $P(y|x)$, the conditional probability of the second sound in the bigram given the first. These probabilities were estimated using add-one smoothing (Chen and Goodman, 1999) from 18,472 citation forms in the TELL database (Inkelas et al., 2000) using the UCI Phonotactic Calculator (Mayer et al., in press). The conditional probabilities assigned to each bigram are shown in Fig. 4. Note that both backness harmony and rounding harmony are reflected in these probabilities: for the most part, harmonic sequences have higher probabilities than disharmonic ones (though other constraints are also apparent, such as a strong dispreference for /ø/ and /o/ in non-initial position).

The UCI Phonotactic Calculator returns log probabilities to avoid numerical underflow. The results in Section 4.7 use these log probabilities rather than the standard probabilities shown in Fig. 4.

4.6.1 Δ in the boolean model

We will test three variants of the boolean model. The first we will call the *harmony* model, based on

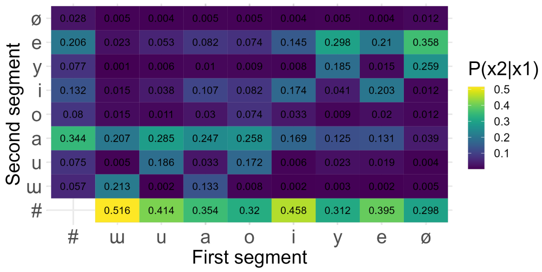

Figure 4: The probability model

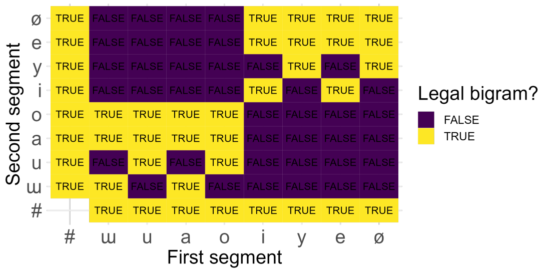

Figure 5: The boolean harmony model

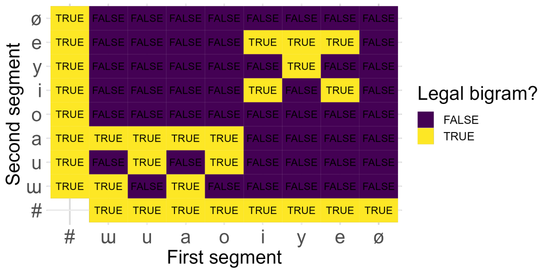

Figure 6: The boolean exception filtering model

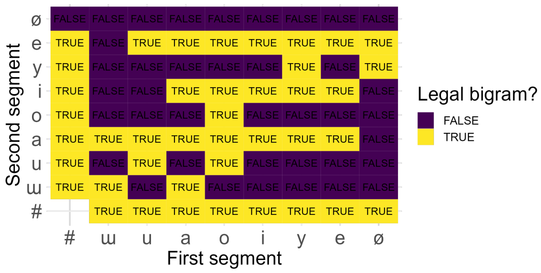

Figure 7: The boolean threshold model

Gorman (2013). Under this model, any bigram that violates either rounding or backness harmony (or both) receives a value of 0 and all other bigrams receive a value of 1. This model is shown in Fig. 5.

The second variant we will call the *exception filtering* model. This is a categorical Turkish phonotactic grammar from Dai (2025), which was learned by a statistical exception filtering process. For reasons of space I will not described the filtering process here, but it results in a more restrictive boolean model that still reflects backness and rounding harmony. This model is shown in Fig. 6.

The third variant we will call the *threshold* model. Under this model, a bigram is legal only if its conditional probability (as defined in the previous section) is above the 40th percentile of all the conditional bigram probabilities. The 40th percentile was opportunistically chosen because it maximized the performance of the model against this data. This is similar to the exception filtering model in that it is derived from frequencies in the lexicon, but it is generally more permissive. The values assigned by this model are shown in Fig. 7.

Gorman (2013) and Kostyszyn and Heinz (2022) also explore models where bigrams are only grammatical if they are attested. Unfortunately, all vowel bigrams are attested in TELL, which means such a model makes no predictions in this case.

4.6.2 Δ in the cost model

We consider only a single variant of the cost model, which uses the same bigram constraints as the harmony model but assigns them integer values instead. Bigrams that violate both backness and rounding harmony have a cost of 2; bigrams that violate one or the other have a cost of 1; and all other bigrams have a cost of 0. The values assigned to bigrams by this model are shown in Fig. 8.

4.7 Results

Each of the five models was used to score the 576 words from the acceptability judgment study. The model scores were correlated against the mean of the normalized acceptability scores for each word collected in the study. Table 2 reports Pearson, Kendall and Spearman correlations (See Albright, 2009, for some discussion of differences between these metrics in the context of phonotactics).

Value type	**Constraints**	r	τ	ρ
Probability	Cond. probs	**0.54**	0.36	**0.50**
Boolean	Threshold	0.46	**0.37**	0.45
Cost	Harmony	0.38	0.30	0.38
Boolean	Harmony	0.38	0.30	0.37
Boolean	Exception	0.36	0.27	0.33

Table 2: Correlations between model scores and mean acceptability judgments.

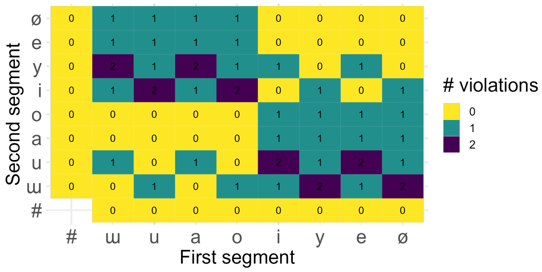

Figure 8: The cost harmony model

These results generally support the probabilistic model as the best approximation of human acceptability judgments. The boolean threshold model comes the closest to matching its performance (and modestly surpasses it according to Kendall's τ). It is important to consider, however, that this model is derived from the conditional probability model: in other words, the best performing categorical model was produced by attending to gradience in the learning data. This is exactly the kind of model argued against by Chomsky (1957), where we "sharpen the blurred edges in the full statistical picture" (p. 17) by designating high probability forms as grammatical and low probability forms as ungrammatical.

Chomsky's objections aside, two natural questions the threshold model must deal with are (a) why the learner should track variability during acquisition only to discard it once the grammar is formed; and (b) how the threshold separating grammatical and ungrammatical structures is set. The learning algorithm in Dai (2025) uses a similar thresholding parameter to determine whether a bigram is exceptional or not. However, Dai finds that the best values of this threshold differ across data sets, and provides no principled way to derive it from the data. In contrast, the conditional bigram model is fit using maximum likelihood estimation, a robust and well-understood learning procedure.

These results favor the use of gradient models for modeling phonotactics. However, in the remainder of the paper I hope to convince you that the similarities between these models outweigh their differences.

5 Reconciling gradient and categorical models

Although these three model types differ in the values they assign to strings, there are many similarities in their basic structure. The boolean, cost, and probability models all assign some value to each segmental bigram (booleans, integers, or probabilities respectively) and aggregate them to get a single value for a string using some binary operation (conjunction, addition, or multiplication respectively). Approaching the models from this perspective, we can abstract away from the specific values and aggregation methods and express them in more mathematically general terms.

Δ maps bigrams to some set of values $\mathcal{T}$:

$$\Delta\colon \Sigma^2 \longrightarrow \mathcal{T}$$

Our `score` function aggregates these values using some binary operator $\owedge$ over $\mathcal{T}$:

$$\text{score}(x_1 \dots x_n) = \bigowedge_{i=1}^{n-1} \Delta(x_i, x_{i+1})$$

The boolean, cost, and probability models described above can be instantiated from this more abstract model by specifying particular values of $\mathcal{T}$ and $\owedge$.

If $\owedge$ is associative and there is an identity element $\top$ in $\mathcal{T}$ such that $a \owedge \top = \top \owedge a = a$, which is the case for each of the set-operation pairs considered here, then $(\mathcal{T}, \owedge)$ forms a mathematical object called a *monoid*. Thinking in monoid-general terms allows us to take the same abstract model and parameterize it with different monoids. This means the same underlying model can compute different quantities, unifying models that appear to do vastly different things on the surface (Goodman, 1999; Eisner, 2003; Chandlee and Heinz, 2017). In other words, we can separate the structure of the model from the values it computes.

In addition to the monoids discussed above, our humble bigram model can actually compute a range of other useful quantities, such as constraint violation profiles using the monoid $(\mathbf{N}^k, +)$, where $\mathbf{N}^k$ is the set of vectors of natural numbers of length k (e.g. Riggle, 2009), or even input SL-2 string transduction (e.g. Chandlee, 2014) using the monoid $(\Sigma^*, \cdot)$, where $\cdot$ is a string concatenation operator.

Most of the models we work with in formal language theory, such as subregular models (Heinz, 2018), finite-state automata, context-free grammars, and so on, can be expressed in these general terms. Although non-deterministic models require an additional operator to combine multiple parses of the same string, a more complex mathematical structure called a *semiring* can be used analogously to monoids for such models.[3]

[3] The probability monoid/semiring is usually defined to

5.1 Monoids in phonology

Why is the idea of monoids useful for us as phonologists? An example comes from the domain of semantics: Giorgolo and Asudeh (2014) apply different semirings to the same underlying semantic model to capture differences between heuristic and mathematical reasoning. They suggest that the underlying structure of both reasoning processes is the same, but that these processes can generate different types of outcomes depending on the context (in this case, how important it is to be precise).

There's perhaps an analogy to be made here with our categorical and gradient models of Turkish. It is clear from the results above and past work on Turkish that vowel harmony is centrally important for both suffix allomorphy and phonotactics (it is striking how much of the variation in participants' responses above can be captured by only attending to the vowels in each word). However, these sensitivities manifest in different ways in each domain. Harmony constraints are essentially categorical when determining suffix allomorphy (it's always [kediler] and never *[kedi-lar]), but these constraints provide only a gradient preference when determining word acceptability.

Even if we choose to treat alternations as essentially categorical and phonotactics as essentially gradient, our categorical and gradient models have more in common than might be evident at first glance. Each of the models we discussed in this paper are TSL-2 grammars: they employ the same types of representations (segments, constraints, etc.); they operate on the vowel tier; they are sensitive only to constraints between adjacent vowels; and they disprefer the same types of structures. The fact that these same representations and dependencies appear to be necessary for modeling both gradient and categorical phenomena suggest that both are governed at least in part by the same underlying linguistic system (Hayes, 2000), and past work has claimed that there is a close connection between the acqusition of alternations and phonotactics (e.g. Hayes, 2004; Chong, 2021; Jun et al., 2025)

assign values from $\mathcal{R}$, with the additional implicit restriction that the assigned values must form a valid probability distribution. There are non-trivial issues that arise in choosing exactly *which* particular values (or *weights*, to use the more technical term) our model should assign, such as normalization in probabilistic models, whether the order of the values is total and monotonic, etc. These considerations are not the focus of this paper.

6 Conclusion

Durvasula (2020) implores us to prioritize categorical models of phonotactics so that we can "focus on what's a possible constraint or rule" and "commit to a specific set of representations." I contend that this is a false dichotomy: constraints and representations in the grammar can be studied independently of the values the grammar assigns. This flexibility allows us to engage with a broader range of empirical phenomena for which categorical or gradient models provide better approximations while still relating these phenomena to the same core linguistic knowledge (Hayes, 2000). Although the results of this study support the position that phonotactic knowledge is best captured using gradient models, we can gain insight into the representations and dependencies in the linguistic grammar by considering both types of models.

Acknowledgments

Thanks to Huteng Dai, Karthik Durvasula, Richard Futrell, Kyle Gorman, Jeff Heinz, Jon Rawski, Megha Sundara, Bernard Tranel, three anonymous reviewers, and the attendees of the 2024 Workshop on Formal Language Theory in Morphology and Phonology at the 2024 Meeting of the Linguistic Society of America for their useful feedback and discussions. Thanks also to Cem Babalik and Defne Bilhan for their assistance in designing the stimuli for the Turkish study. This work was supported by NSF Award #2214017.

References

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

Alexander L Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52:388–407.

Todd M Bailey and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4):568–591.

Jane Chandlee. 2014. *Strictly local phonological processes*. University of Delaware.

Jane Chandlee and Jeffrey Heinz. 2017. Computational phonology. In *Oxford Research Encyclopedia of Linguistics*.

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

Eleanor Chodroff and Colin Wilson. 2014. Phonetic vs. phonological factors in coronal-to-dorsal perceptual assimilation. Paper presented at LabPhon 14: the 14th Conference on Laboratory Phonology, Tokyo.

Noam Chomsky. 1957. *Syntactic structures*. Walter de Gruyter.

Noam Chomsky and Morris Halle. 1968. *The sound pattern of English*. Harper & Row, New York.

Adam J Chong. 2021. The effect of phonotactics on alternation learning. *Language*, 97(2):213–244.

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.

Huteng Dai. 2025. An exception-filtering approach to phonotactic learning. *Phonology*, 42:e5.

Huteng Dai, Connor Mayer, and Richard Futrell. 2023. Rethinking representations: A log-bilinear model of phonotactics. *Proceedings of the Society for Computation in Linguistics*, 6.

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

Emmanuel Dupoux, Erika Parlato, Sonia Frota, Yuki Hirose, and Sharon Peperkamp. 2011. Where do illusory vowels come from? *Journal of memory and language*, 64(3):199–210.

Karthik Durvasula. 2020. O gradience, whence do you come? Keynote presentation at the 2020 Annual Meeting on Phonology.

Jan Edwards, Mary E Beckman, and Benjamin Munson. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition.

Jason Eisner. 2003. Simpler and more general minimization for weighted finite-state automata. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 64–71.

G. Giorgolo and A. Asudeh. 2014. One semiring to rule them all. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, pages 208–226. Cognitive Science Society, Québec City.

Matthew Goldrick and Meredith Larson. 2008. Phonotactic probability influences speech production. *Cognition*, 107(3):1155–1164.

Sharon Goldwater and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Östen Dahl, editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120. Stockholm University, Department of Linguistics, Stockholm.

Josh Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25(4):573–605.

Kyle Gorman. 2013. *Generative phonotactics*. Ph.D. thesis, University of Pennsylvania.

Bruce Hayes. 2000. *Gradient well-formedness in Optimality Theory*, pages 88–120. Oxford University Press.

Bruce Hayes. 2004. Phonological acquisition in optimality theory: The early stages. *Constraints in Phonological Acquisition/Cambridge University Press*.

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

Jeffrey Heinz. 2018. The computational nature of phonological generalizations. *Phonological typology, phonetics and phonology*, 23:126–195.

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.

Sharon Inkelas, Aylin Küntay, Orhan Orgun, and Ronald Sprouse. 2000. Turkish electronic living lexicon (TELL). *Turkic Languages*, 4:253–275.

Jongho Jun, Hanyoung Byun, Seon Park, and Yoona Yee. 2025. How tight is the link between alternations and phonotactics? *Phonology*, 42:e3.

Jimin Kahng and Karthik Durvasula. 2023. Can you judge what you don't hear? Perception as a source of gradient wordlikeness judgments. *Glossa*, 8(1).

Kalina Kostyszyn and Jeffrey Heinz. 2022. Categorical account of gradient acceptability of word-initial Polish onsets. In *Proceedings of AMP 2021*.

Andrey Andreyevich Markov. 1913. An example of statistical investigation in the text of 'Eugene Onyegin' illustrating coupling of 'tests' in chains. *Proceedings of the Academy of Sciences of St. Petersburg*, 7:153–162.

Sven L Mattys, Peter W Jusczyk, Paul A Luce, and James L Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38(4):465–494.

Connor Mayer. 2021. Capturing gradience in long-distance phonology using probabilistic tier-based strictly local grammars. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 39–50, Online. Association for Computational Linguistics.

Connor Mayer, Arya Kondur, and Megha Sundara. in press. The UCI Phonotactic Calculator: An online tool for computing phonotactic metrics. *Behavior Research Methods*.

Connor Mayer, Arya Kondur, and Megha Sundara. under revision. The UCI Phonotactic Calculator: An online tool for computing phonotactic metrics. *Behavior Research Methods*.

Connor Mayer, Adeline Tan, and Kie Ross Zuraw. 2024. Introducing maxent. ot: an R package for Maximum Entropy constraint grammars. *Phonological Data and Analysis*, 6(4):1–44.

James M McQueen. 1998. Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, 39(1):21–46.

Karima Mersad and Thierry Nazzi. 2011. Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory & Cognition*, 39:1085–1093.

Dennis Norris and James M. McQueen. 2008. Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115:357–395.

Alan Prince and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Jason Riggle. 2009. Violation semirings in optimality theory. *Research on Language and Computation*, 7:1–12.

Robert Scholes. 1966. *Phonotactic grammaticality*. Mouton, The Hague.

Carson Schütze. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. The University of Chicago.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656.

Jeremy Steffman and Megha Sundara. 2023. Short-term exposure alters adult listeners' perception of segmental phonotactics. *JASA Express Letters*, 3(12).

Jeremy Steffman and Megha Sundara. 2024. Disentangling the role of biphone probability from neighborhood density in the perception of nonwords. *Language and Speech*, 67(1):166–202.

Holly L Storkel. 2001. Learning new words. *Journal of Speech, Language, and Hearing Research*, 44(6):1321–1337.

Conrad F Taylor and George Houghton. 2005. Learning artificial phonotactic constraints: time course, durability, and relationship to natural constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6):1398.

Michael S Vitevitch and Paul A Luce. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of memory and language*, 40(3):374–408.

Jill A Warker. 2013. Investigating the retention and time course of phonotactic constraint learning from production experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(1):96.

Jill A Warker and Gary S Dell. 2006. Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2):387.

Jill A Warker and Gary S Dell. 2015. New phonotactic constraints learned implicitly by producing syllable strings generalize to the production of new syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(6):1902.

K.E. Zimmer. 1969. Psychological correlates of some Turkish morpheme structure conditions. *Language*, pages 309–321.