One (semi)ring to rule them all: Reconciling categorical and gradient models of phonotactics

LSA Session on Formal Language Theory in Morphology and Phonology

Connor Mayer
January 6th, 2024

Department of Language Science
University of California, Irvine
Question: Are phonotactic grammars categorical or gradient?

Answer: It depends on which semiring you use!
Two points I want to make:

1. Gradient phonotactic models account for new data from a Turkish acceptability judgment task better than categorical models.

2. This distinction turns out to be somewhat superficial if we think of models from a semiring-general perspective.
What is phonotactics?
What is phonotactics?

The legal ways in which sounds can be sequenced into words.

This is (mostly) learned and language-specific:

- /stik/ would be an ok English word; not a good Spanish word
- /tʃkɔwɛntʃ/ is a fine Polish word; not a good English word
Phonotactics is gradient

A typical source of data is to ask speakers for acceptability judgments:

- “on a scale of 1-7, how likely is ‘steek’ to become an English word?”
- “would ‘steek’ be a better English word than ‘chknonch’?”
- “could ‘steek’ be an English word?”

What do we mean by gradience?

poik
lvag
kip
What do we mean by gradience?

Ivag ≪ poik ≪ kip
Where does this gradience come from?

Gradience in acceptability judgments can arise from performance factors such as misperception [e.g. Kahng and Durvasula, 2023].

However, the gradience observed in phonotactic acceptability judgments is largely predictable from “soft” versions of the same constraints that govern other phonological processes [Hayes, 2000].

Typical modeling approach is to use a grammar that produces a gradient output.

- Often based on statistical frequencies in the lexicon.
Implementing a gradient phonotactic grammar

Our phonotactic grammars consist of a score function that assigns values to words.

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

Such a model can represent gradient acceptability judgments:

\[ \text{score}(\text{lvag}) = 0.01 < \text{score}(\text{poik}) = 0.2 < \text{score}(\text{kip}) = 0.4 \]
Is phonotactics categorical?
Is phonotactics categorical?

Gorman [2013] argues that we have been premature in assuming the phonotactic grammar computes gradient outputs.

- **Proposal**: grammar is **categorical** and gradience comes from other sources.
- A categorical grammar labels words as either grammatical or ungrammatical

In particular, he claims that **categorical models do as well as or better than gradient models** in predicting phonotactic phenomena.
Past work on categorical grammars

Categorical models have been claimed to better predict:

- English onset acceptability [Gorman, 2013, Durvasula, 2020, Dai, accepted]
- Polish onset acceptability [Kostyszyn and Heinz, 2022, Dai, accepted]
- Turkish vowel distributions [Gorman, 2013, Dai, accepted]
- English medial cluster distributions [Gorman, 2013]
Limitations of previous work

1. Use a very small number of data sets, almost all about consonant clusters
2. Authors have different definitions of “categorical”
3. The gradient model used in (almost) all cases is the *UCLA Phonotactic Learner*
Limitation 2: Defining categorical

Some “categorical” models are in fact gradient [Durvasula, 2020, Kostyszyn and Heinz, 2022].

- Words receive an integer score corresponding to number of constraint violations
- “Categorical” in these models means all constraint violations are penalized equally

These models can represent a situation where $lvag \ll poik \ll kip$.

For the sake of time I’m going to ignore these models.
Limitation 2: Defining categorical

Other proposed categorical models are truly categorical [Gorman, 2013, Kostyszyn and Heinz, 2022, Dai, accepted]

- Words are grammatical or not
- I’ll refer to this as a \textit{boolean} model of phonotactics

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

These models \textit{cannot} represent a situation where lvag \ll poik \ll kip

\textbf{We’ll adopt this definition of categorical.}
The UCLA Phonotactic Learner has become the poster boy for gradient phonotactics [Hayes and Wilson, 2008].

- But it also has to learn constraints from the data!
- Its performance is sensitive to how it is parameterized.
- Do categorical models outperform it because it is gradient? Because of its constraint selection process? Because it has been run with sub-optimal hyperparameters?
Let’s compare the performance of two proposed categorical boolean models of Turkish vowel phonotactics against a simple probabilistic bigram model with a similar structure.

We’ll evaluate how these models predict new experimental data from a Turkish nonce word acceptability judgment task.
A new dataset of Turkish acceptability judgments
<table>
<thead>
<tr>
<th></th>
<th>[−back]</th>
<th></th>
<th>[−round]</th>
<th>[+round]</th>
<th>[−round]</th>
<th>[+round]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[−high]</td>
<td>i</td>
<td>y</td>
<td>ω</td>
<td>u</td>
<td>u</td>
<td></td>
</tr>
<tr>
<td>[−high]</td>
<td>e</td>
<td>φ</td>
<td>a</td>
<td>o</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Constraints on Turkish vowels

**Backness harmony:** *[a]back [−a]back

- A vowel must agree in backness with the preceding vowel.

**Rounding harmony:** *[a]round [−a]round, +high]

- A high vowel must agree in roundness with the preceding vowel.

These constraints govern suffix allomorphy, but their effect is also detectable in the lexicon and in acceptability judgment tasks [Zimmer, 1969].
Our data

The data we’ll look at are acceptability judgments from a large, online study.

- **Participants**: 90 native Turkish speakers (38F; mostly age 25-35) recruited on Prolific

- **Task**: Wug word acceptability judgments
Stimuli

**Stimuli**: 596 wug words with CVCVC shape

- Nine words for each unique pair of vowels (8 × 8 total pairs)
- Probability of consonants controlled for within vowel groups
- Synthesized to speech using Google Cloud
- Words and recordings vetted by two native Turkish speakers
Experiment task

Deney

Kalan süre

beyop

lvag  caçör  matan
Each participant rated 192 tokens after training and attention checks: 17,280 tokens.

Responses are normalized to z-scores within participant

- Controls for differences in mean and spread between participants
We’ll test three simple models that have similar structures:

<table>
<thead>
<tr>
<th>Value type</th>
<th>Constraint values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>Conditional probabilities</td>
</tr>
<tr>
<td>Boolean</td>
<td>Harmony [Gorman, 2013]</td>
</tr>
<tr>
<td>Boolean</td>
<td>Exception filtering [Dai, accepted]</td>
</tr>
</tbody>
</table>
General model structure

All the models are TSL-2 grammars that operate on the vowel tier

- Informally, we **ignore consonants** and assign scores based on **vowel bigrams**

- Constraints can reference start and end symbols ⌊ and ⌋
Scoring bigrams

Each model type has a $\Delta$ function that assigns a value to a bigram.

**Boolean model**

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

**Probability model**

$$\Delta_p : \Sigma^2 \rightarrow [0, 1]$$
Scoring words

**Boolean model:** words are assigned 1 if they contain only legal bigrams, 0 otherwise

\[
\text{bigram\_score}(x_1, \ldots, x_n) = \bigwedge_{i=1}^{n-1} \Delta_b(x_i, x_{i+1})
\]

**Probability model:** words are assigned the product of the probability of each bigram.

\[
\text{probability\_score}(x_1, \ldots, x_n) = \prod_{i=1}^{n-1} \Delta_p(x_i, x_{i+1})
\]
An example

Boolean model

\[
\text{boolean\_score([oi])} = \Delta_b(\otimes o) \land \Delta_b(oi) \land \Delta_b(i\times) \\
= 1 \land 0 \land 1 \\
= 0
\]

Probabilistic model

\[
\text{probability\_score([oi])} = \Delta_p(\otimes o) \times \Delta_p(oi) \times \Delta_p(i\times) \\
= 0.08 \times 0.107 \times 0.458 \\
= 0.0004
\]
Choosing our values

How do we define $\Delta$ for each model?
The probability model uses Laplace-smoothed conditional probabilities derived from 18,472 citation forms in the TELL database [Inkelas et al., 2000].

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>w</th>
<th>u</th>
<th>a</th>
<th>o</th>
<th>i</th>
<th>y</th>
<th>e</th>
<th>ø</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td>0.028</td>
<td>0.004</td>
<td>0.004</td>
<td>0.005</td>
<td>0.004</td>
<td>0.004</td>
<td>0.005</td>
<td>0.004</td>
<td>0.011</td>
</tr>
<tr>
<td>e</td>
<td>0.206</td>
<td>0.023</td>
<td>0.052</td>
<td>0.082</td>
<td>0.073</td>
<td>0.145</td>
<td>0.298</td>
<td>0.21</td>
<td>0.361</td>
</tr>
<tr>
<td>y</td>
<td>0.077</td>
<td>0.001</td>
<td>0.006</td>
<td>0.01</td>
<td>0.009</td>
<td>0.008</td>
<td>0.185</td>
<td>0.015</td>
<td>0.261</td>
</tr>
<tr>
<td>i</td>
<td>0.132</td>
<td>0.015</td>
<td>0.038</td>
<td>0.107</td>
<td>0.081</td>
<td>0.174</td>
<td>0.041</td>
<td>0.203</td>
<td>0.011</td>
</tr>
<tr>
<td>o</td>
<td>0.08</td>
<td>0.015</td>
<td>0.011</td>
<td>0.03</td>
<td>0.074</td>
<td>0.033</td>
<td>0.008</td>
<td>0.02</td>
<td>0.011</td>
</tr>
<tr>
<td>a</td>
<td>0.344</td>
<td>0.207</td>
<td>0.286</td>
<td>0.247</td>
<td>0.258</td>
<td>0.17</td>
<td>0.125</td>
<td>0.131</td>
<td>0.038</td>
</tr>
<tr>
<td>u</td>
<td>0.075</td>
<td>0.005</td>
<td>0.186</td>
<td>0.032</td>
<td>0.172</td>
<td>0.006</td>
<td>0.022</td>
<td>0.019</td>
<td>0.003</td>
</tr>
<tr>
<td>w</td>
<td>0.057</td>
<td>0.213</td>
<td>0.002</td>
<td>0.133</td>
<td>0.007</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>#</td>
<td>0.516</td>
<td>0.415</td>
<td>0.354</td>
<td>0.32</td>
<td>0.458</td>
<td>0.313</td>
<td>0.396</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
Words are grammatical if they satisfy both rounding and backness harmony.

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>u</th>
<th>a</th>
<th>o</th>
<th>i</th>
<th>y</th>
<th>e</th>
<th>ø</th>
</tr>
</thead>
<tbody>
<tr>
<td>ø</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>eyi</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
</tr>
<tr>
<td>eyioa</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
</tr>
<tr>
<td>eyiouw</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>eyiouw</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>eyiouw</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
<tr>
<td>eyiouw</td>
<td>TRUE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>TRUE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

Legal bigram?

- **FALSE**
- **TRUE**
Categorical Turkish phonotactic grammar from Dai [accepted] learned via an exception filtering process.
Results

Let’s look at correlations between model score and mean acceptability judgment.

<table>
<thead>
<tr>
<th>Value type</th>
<th>Constraint set</th>
<th>$r$</th>
<th>$\tau$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>Conditional probabilities</td>
<td>0.558</td>
<td>0.375</td>
<td>0.527</td>
</tr>
<tr>
<td>Boolean</td>
<td>Harmony [Gorman, 2013]</td>
<td>0.371</td>
<td>0.303</td>
<td>0.369</td>
</tr>
<tr>
<td>Boolean</td>
<td>Exception filtering [Dai, accepted]</td>
<td>0.360</td>
<td>0.286</td>
<td>0.348</td>
</tr>
</tbody>
</table>

The simple probabilistic model substantially outperforms the other models.
Reconciling categorical and gradient models using semirings
The reconciliation begins
Probabilistic and boolean TSL-2 models differ:

- **Boolean**: Assigns booleans to segmental bigrams, combines them using $\land$.
- **Probabilistic**: Assigns probabilities to segmental bigrams, combines using $\oplus$.

But the basic structure of each model is the same:

- We assign some **value** to each segmental bigram
- We **aggregate** those values to get a score for the word
Commonalities between categorical and structural models

We can abstract away from specific values/aggregators:

$$\Delta: \Sigma^2 \rightarrow \mathcal{R}$$

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

where $\mathcal{R}$ is some set of values and $\bigwedge$ is some binary operator over $\mathcal{R}$. 
We can make these simple models compute even more interesting quantities!

<table>
<thead>
<tr>
<th>What does it compute?</th>
<th>$\mathcal{R}$</th>
<th>$\bigwedge$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boolean scores</td>
<td>${0, 1}$</td>
<td>$\land$</td>
</tr>
<tr>
<td>[Gorman, 2013, Kostyszyn and Heinz, 2022, Dai, accepted]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilities</td>
<td>$[0, 1]$</td>
<td>$\times$</td>
</tr>
<tr>
<td>Integer scores</td>
<td>$\mathbb{N}$</td>
<td>$+$</td>
</tr>
<tr>
<td>[Durvasula, 2020, Kostyszyn and Heinz, 2022]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constraint violation profiles</td>
<td>$\mathbb{N}^k$</td>
<td>$+$</td>
</tr>
<tr>
<td>Left SL-2 string transduction</td>
<td>$\Sigma^*$</td>
<td>$+$</td>
</tr>
</tbody>
</table>
What's going on here?

This definition of our TSL-2 models is in **semiring-general terms**

\[ \Delta : \Sigma^2 \rightarrow \mathcal{R} \]

\[ \text{score}(x_1 \ldots x_n) = \bigwedge_{i=1}^{n-1} \Delta(x_i, x_{i+1}) \]

We can parameterize our model with different semirings that provide implementations of \( \mathcal{R} \) and \( \bigwedge \).
A semiring is an algebraic structure.

**Monoid**: a set $\mathcal{R}$ closed under a binary relation $\sqcap$ such that:

- $\sqcap$ is associative
- There’s an identity element $\top$ in $\mathcal{R}$ such that $a \sqcap \top = \top \sqcap a = a$
A **semiring** consists of a pair of monoids

- \((R, \bigwedge)\) with identity element \(\top\)
- \((R, \bigvee)\) with identity element \(\bot\)

such that:

- \(\bigwedge\) distributes over \(\bigvee\)
- \(x \bigwedge \bot = \bot \bigwedge x = \bot\)
Why are semirings interesting?

The models we work with in FLT (TSL, FSA, CFG, etc.) can be expressed in semiring-general terms.

- In terms of $R$, $\&, \|$, rather than specific values and operators
- (TSL-2 models don’t use $\|$ but it’s important for more complex models)

Different semirings allow the same underlying model to compute different quantities.

- Unifies superficially different models [Goodman, 1999].

We can separate the structure of the model from the values it computes.
Why is this useful for us as phonologists?

Semirings allow us to relate the grammar to different domains or contexts

- Giorgolo and Asudeh [2014] apply different semirings to the same underlying semantic model to capture differences in heuristic vs. mathematical reasoning.

Connecting the grammar to different domains

There’s perhaps an analogy to be made to Turkish.

- Harmony is essentially categorical when determining suffix allomorphy

  ‘cat-PL‘  kedi-ler ✓  kedi-lar ✗

- Harmony is a gradient preference when determining word acceptability

- **But both sensitive to the same configurations!**
Connecting the grammar to different domains

Regardless of semiring, both the categorical and probabilistic grammars we saw here

- are sensitive only to bigram constraints
- use segmental representations
- operate on the vowel tier

These are segmental TSL-2 grammars, regardless of the values they assign.

The same applies to other representations or grammars.
Durvasula [2020] closes with a plea to abandon gradience and adopt categorical grammars so we can

• “focus on what’s a possible constraint or rule”; and
• “commit to a specific set of representations”
This is a false dichotomy.

- Constraints and representations in the grammar can be studied independently of the values the grammar assigns.

- Insight into the structure of the grammar can come from both gradient and categorical analyses!

- This flexibility allows our models to engage with a broader range of empirical phenomena.
Thanks to Huteng Dai, Jon Rawski, Megha Sundara, and Richard Futrell for many interesting discussions, and to my Turkish consultants Cem Babalik and Defne Bilhan.


Huteng Dai. An exception-filtering approach to phonotactic learning, accepted.


Stimuli structure

- Log unigram probability
- Log smoothed bigram probability
Cost semiring

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>w</th>
<th>u</th>
<th>a</th>
<th>o</th>
<th>i</th>
<th>y</th>
<th>e</th>
<th>ø</th>
</tr>
</thead>
<tbody>
<tr>
<td>ø</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>i</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>u</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Constraint value:
- 0
- 1
- 2
### Results

<table>
<thead>
<tr>
<th>Value type</th>
<th>Constraint set</th>
<th>$r$</th>
<th>$\tau$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>Conditional probabilities</td>
<td>0.558</td>
<td>0.375</td>
<td>0.527</td>
</tr>
<tr>
<td>Boolean</td>
<td>Cost</td>
<td>-0.379</td>
<td>-0.305</td>
<td>-0.386</td>
</tr>
<tr>
<td></td>
<td>[Durvasula, 2020, Kostyszyn and Heinz, 2022]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boolean</td>
<td>Harmony [Gorman, 2013]</td>
<td>0.371</td>
<td>0.303</td>
<td>0.369</td>
</tr>
<tr>
<td>Boolean</td>
<td>Exception filtering [Dai, accepted]</td>
<td>0.360</td>
<td>0.286</td>
<td>0.348</td>
</tr>
</tbody>
</table>