# A Neo-Trubetzkoyan approach to phonotactic learning in the presence of exceptions*

Huteng Dai

Jan 30, 2023

**Abstract**

Phonotactic learning is a crucial aspect of phonological acquisition and plays a significant role in computational research in phonology. This paper proposes a "categorical grammar + statistical criterion" approach that infers categorical phonotactic grammars from realistic corpora by iteratively filtering out ungrammatical sequences based on type frequency information. The learned grammars are shown to capture categorical phonotactic patterns and accurately fit gradient acceptability judgments. The learned grammar captures distributional generalizations in English and Turkish based on both corpora and experimental data. I argue that this approach is at least as good as, and appears in several cases superior to, the "Probabilistic grammar + Probabilistic inference" methods in handling exceptions.

**Keywords**: phonotactics; phonological learning, indirect negative evidence, type frequency, onsets, gradient acceptability, Turkish

## 1 Introduction

This paper proposes a learning model that simulates how humans extract a *categorical* grammar of phonotactics in the presence of *exceptions*. While maintaining a categorical grammar, this proposal incorporates *frequency information* to identify potential constraints. The proposed learner successfully induces categorical phonotactic generalizations of English and Turkish from large-scale naturalistic corpora containing exceptions. The predicted judgments of nonce words by the learner are significantly correlated with the categorical and gradient acceptability of nonce words (Daland et al., 2011). The current study not only provides an alternative to probabilistic approaches but also sheds light on the fundamental question in language acquisition: how do human brains, as biological computing systems, detect patterns in natural language corpora that may contain misleading data (Angluin and Laird, 1988; Perkins et al., 2022)? Moreover, although

*This manuscript has been submitted to *Phonology*. Comments and suggestions are welcome (Email: hutengdai@gmail.com). I thank Colin Wilson, Adam Jardine, Yang Wang, Adam McCollum, Jeff Heinz, and Jon Rawski for their suggestions. All mistakes are mine.

the proposal is based on categorical grammar, this paper contributes to the studies of probabilistic models by analyzing how probabilistic learning models behave in the presence of exceptions.

A categorical grammar of any linguistic domain, especially syntax and phonology, assigns discrete values to the possible *representations*. The representation in the domain of phonotactics is sound sequences (phonological words). In particular, a representation is either grammatical or ungrammatical in binary categorical grammars. In contrast, a probabilistic grammar assigns continuous values of probabilities to representations and captures the "degrees", rather than the "categories", of grammaticality. Previous research on phonological learning has focused on approaches based on probabilistic grammars, such as the Maximum Entropy learner (henceforth "MaxEnt learner") and its variants (Hayes and Wilson, 2008; Wilson and Gallagher, 2018; Gouskova and Gallagher, 2020; Hughto et al., 2019). However, probabilistic approaches are not necessarily superior in handling exceptions, as previous studies claimed (Wilson and Gallagher, 2018). Although a probabilistic grammar predicts a distribution that naturally matches the surface frequency of sequences, it also under-penalizes ungrammatical sequences from the learning sample (discussed §6.2). This issue is naturally resolved in a categorical grammar.

This article is structured as follows: §2 outlines the background of the research problem in this paper; §3 introduces my proposal; §4 and §5 apply my proposal to the English and Turkish corpora and compare my proposal with the MaxEnt learner; §6 discusses open questions and future directions.

# 2 Background

This section introduces the learning problem in the presence of exceptions and the basics of formal methods.

## 2.1 Phonotactic learning in the presence of exceptions

The distinction between *attestedness* and *grammaticality* is crucial for understanding exceptionality. As illustrated in Table 1, attestedness indicates whether a sequence of phonological representations occurs in the learning data ("Does this sequence exist?" in perceptual experiments). For example, *brick* and *sphere* are both attested as they appear in English, while *blick* and *bnick* are not. In contrast, grammaticality indicates the well-formedness and productivity of phonological representations for a native speaker (induced from prompts such as "How acceptable is this sequence?" in perceptual experiments). For instance, *sphere* in English is ungrammatical because the [*sf] sequence is neither well-formed nor productive—native speakers consider a nonce word with [*sf] ungrammatical (or "exotic" per Hayes and Wilson, 2008, P395). Speakers' acceptability judgments, gradient or categorical, reflect the grammaticality: grammatical sequences are more acceptable than ungrammatical sequences (Albright, 2009; Gorman, 2013; Lau et al., 2017).

| | grammatical | ungrammatical |
|---|---|---|
| **attested** | *brick* | *sphere* |
| **unattested** | *blick* | *bnick* |

Table 1: The distinction between attestedness and grammaticality (adapted from Hyman, 1975)

Human learners cannot simply memorize infinite combinations of speech sounds. Instead, they must internalize abstract phonotactic constraints from the learning sample to differentiate acceptable and unacceptable sequences of sounds. For example, although both are nonexistent words, English speakers consider *blick* more acceptable than *bnick*. *bnick* contains an ungrammatical sequence [bn], and **bn* is an abstract phonotactic constraint that causes the ungrammaticality of *bn*. Speakers usually use this knowledge unconsciously to create new words in their language.

Figures 1 and 2 illustrate the problem of phonotactic learning: the dots indicate a subset of possible English onsets; the darker dots represent attested onsets, e.g. [sf], and lighter dots represent unattested ones, e.g. [vw]. learners are only exposed to *positive evidence* from the attested sample, meaning they do not have access to the grammaticality label for each onset that is only available in negative evidence. The learning problem, as shown in Figure 2, is to find the target grammar that successfully distinguishes grammatical and ungrammatical sequences for both the attested (darker dots) and unattested (lighter dots) data. As indicated by the curve, the target grammar must predict the grammaticality of exceptions.

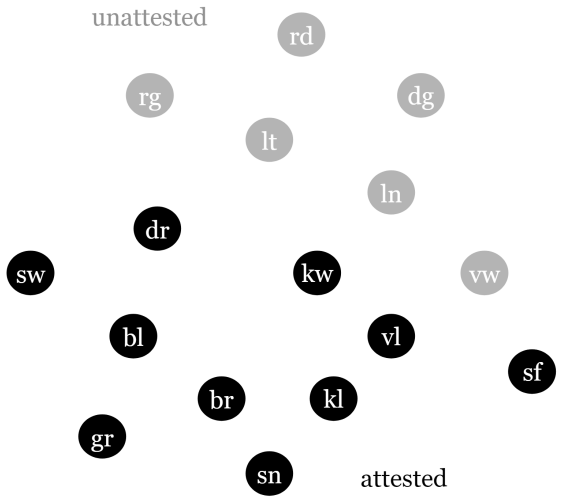



Figure 1: Learning sample

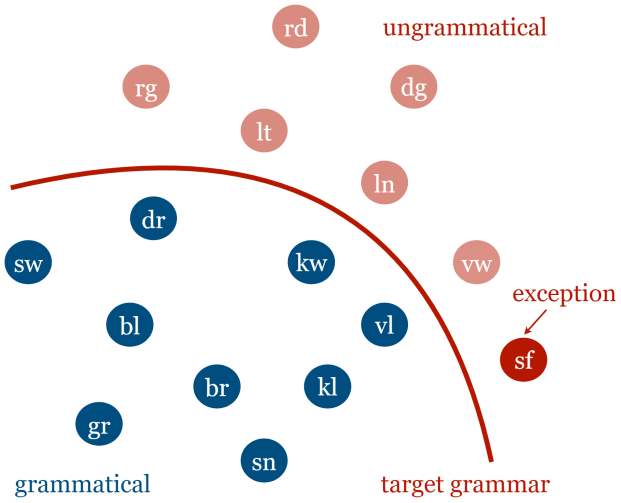



Figure 2: Target grammar that distinguishes grammatical from ungrammatical words; dark vs. light color contrast indicates attested vs. unattested forms

However, the mismatch between attestedness and grammaticality gives rise to *exceptions* (attested but ungrammatical) and *accidental gaps* (unattested but grammatical), which significantly challenges phonological learning. Accidental gaps have been emphasized in previous literature (Chomsky and Halle, 1965; Hayes and Wilson, 2008; Wilson and Gallagher, 2018), while the proper treatment of exceptions remains understudied. Most previous work on exceptions focuses on morphophonological processes (Zuraw, 2000; Linzen et al., 2013; Moore-Cantwell and Pater, 2016; Yang, 2016; Hughto et al., 2019), which is different from phonotactic learning in terms of data (UR-SR pairs vs. SRs) and grammar (constraint ranking/rule ordering vs. constraint sets). The current paper focuses on exceptions in *phonotactic* learning and leaves accidental gaps for future studies.

The challenge of *exceptions*, defined as ungrammatical yet attested sequences, results from the assumptions of a positive evidence-only setting and categorical grammaticality. Without negative evidence (input labels), if learners assume every attested sound sequence is grammatical, they will misgeneralize exceptions as grammatical. The learning problem in the English example is: how do native speakers learn that [zw], [sf], and [pw] are ungrammatical (*exotic* per Hayes and Wilson, 2008) when these sequences appear in attested words such as *Zwieback*, *sphere*, and *Puerto Rico*.

## 2.2 Indirect negative evidence

The learning problem in the presence of exceptions is challenging because intrusions of ungrammatical sequences are misleading in a positive evidence-only setting. Computationally, when noisy input is present, a Gold-style learning model exposed to only positive evidence (Gold, 1967) cannot learn the target grammar in most classes of formal languages (Osherson et al., 1986), especially linguistically interesting classes such as Strictly 2-Local languages described by bigram

constraints.

The solution to this problem lies in *indirect negative evidence* from frequency information (Clark and Lappin, 2009, 2010; Regier and Gahl, 2004; Pearl and Lidz, 2009; Pearl and Mis, 2016), in particular type frequency (Pierrehumbert, 2001; Albright and Hayes, 2003; Hayes and Londe, 2006; Hayes and Wilson, 2008; Albright, 2009). The learner can utilize the statistical knowledge to conclude that the occurrences of this sequence are ungrammatical exceptions. For example, if an attested sequence *sf only appears in a few English words (low type frequency) in the dataset, then the learner will use this information to infer that *sf is ungrammatical.

## 2.3 Probabilistic vs categorical approaches

Although probabilistic approaches are more common in the literature, indirect negative evidence can also be incorporated into categorical approaches. This section first introduces probabilistic approaches, then discusses categorical approaches.

### 2.3.1 Probabilistic approaches

The common way of using indirect negative evidence is to replace the categorical grammar with a probabilistic grammar i.e. *language model* (or *phonotactic model* in terms of computational modeling). A standard paradigm is to model how a phonotactic model assigns word likelihood ("production"), then models the statistical learning via Maximum Likelihood Estimation (Albright, 2009) or Maximum Entropy (Hayes and Wilson, 2008). The MaxEnt grammar predicts a probabilistic distribution from the MaxEnt values $P^* = \exp(-h)$, where $h$ is the harmony score accumulated from the weights of violated constraints, as defined below:

$$h(s) = \sum_i w_i \times C_i(s) \tag{1}$$

And the probabilities Pr are obtained by normalizing $P^*$. As the learner maximizes the likelihood of the observed data, the learned probabilistic grammar approximates a distribution that matches the surface frequency. As illustrated in Table 2, a hypothesis MaxEnt grammar will assign a higher weight to *sf than *bɹ if [sf] is less frequent than [bɹ]. A nonce word [sfɪn] receives a lower probability than [bɹɪki] due to the higher weight on the constraint *sf in the predicted probabilistic distribution.

| | $w = 2$ | $w = 1$ | | | |
|---|---|---|---|---|---|
| | *sf | *bɹ | $h$ | $P^* = \exp(-h)$ | Pr |
| sfɪn | 1 | 0 | 2 | 0.135 | 0.245 |
| bɹɪki | 0 | 1 | 1 | 0.368 | 0.665 |
| bɹɪkibɹɪkibɹɪki | 0 | 3 | 3 | 0.050 | 0.090 |

Table 2: Nonce words in a hypothesis MaxEnt grammar

Probabilistic grammars conflate grammaticality into one spectrum of real-valued probabilities, eliminating the boundary between grammatical and ungrammatical sequences, including

exceptions. Ideally, "ungrammatical" sequences should receive lower probabilities than "grammatical" sequences in a probabilistic grammar. However, a long word considered grammatical in categorical grammar might instead receive a lower probability than a shorter ungrammatical exception in a probabilistic grammar, because probabilistic grammars intrinsically penalize long words as they often accumulate more constraint violations (Heinz and Idsardi, 2017). For example, although the nonce word [bɹɪkibɹɪkibɹɪki] is simply a reduplication of [bɹɪki], it receives a lower score than [sfɪn], because it violates the lower weight constraint *bɹ three times. Consequently, it is impossible to find the exact score threshold that separates grammatical and ungrammatical sequences in a probabilistic grammar. A learner based on probabilistic grammar cannot filter out exceptions during the learning procedure. As discussed in §6.2, the MaxEnt learner will internalize exceptions as grammatical words, which eventually affects its capability to handle exceptionful samples.

As illustrated in Table 3, one potential solution to this problem is to assign an extremely high weight for exceptions such as $w = 200$ for *sf.

| | $w = 200$ | $w = 1$ | | | |
|---|---|---|---|---|---|
| | *sf | *bɹ | $h$ | $P^* = \exp(-h)$ | Pr |
| sfɪn | 1 | 0 | 200 | $1.384 \times 10^{-87}$ | $3.313 \times 10^{-87}$ |
| bɹɪki | 0 | 1 | 1 | 0.368 | 0.881 |
| bɹɪkibɹɪkibɹɪki | 0 | 3 | 3 | 0.050 | 0.119 |

Table 3: The hypothetical learned MaxEnt grammar where *sf receives $w = 200$

This solution penalizes ungrammatical sequences [sfn] so that it receives a lower score than a long word [bɹɪkibɹɪkibɹɪki]. However, a grammatical word might be long enough to receive a lower score than the exception.

Another solution to the problem of exceptions in MaxEnt learners is to use lexical indices and lexically specified constraints (Pater, 2000; Moore-Cantwell and Pater, 2016; Hughto et al., 2019). Instead of excluding exceptions from the grammar, this approach adds an additional mechanism for indexed exceptions in the grammar. For example, a lexically specific constraint *$\text{sf}_i$ will penalize the sequence [sf] unless it appears in an indexed lexical exception *sphere*. However, this approach cannot be extended to nonce words where prespecified lexical indices and additional constraints are not available.

### 2.3.2 The power of ignoring: categorical approaches

Categorical grammars maintain a clear distinction between grammatical and ungrammatical sequences, including exceptions, and avoid issues in probabilistic grammar mentioned above. In addition, the current article argues that indirect negative evidence can also be used in categorical approaches to handle exceptions.

Previous categorical models of phonotactic learning, although performing well in several datasets (Gorman, 2013; Kostyszyn and Heinz, 2022), departed from the problem of exceptions. Gouskova and Gallagher (2020) stated that:

> "In contrast to our approach, Heinz (2010), Jardine (2016), and Jardine and Heinz (2016) characterize nonlocal phonology as an idealized problem of searching for unattested substrings. Their learners memorize attested precedence relations between segments and induce constraints against those sequences that they have not encountered. One of the problems with this approach is that it can reify *accidental gaps* to the level of categorical phonotactic constraints, whereas stochastic patterns with *exceptions* will stymie it (Wilson and Gallagher, 2018)."

Although they correctly pointed out that the baseline categorical learning models cannot generalize to accidental gaps or handle exceptions, Gouskova and Gallagher (2020) assumed that probabilistic models *inherently* solve the issue of exceptions, which is not justified in any previous literature. Furthermore, Heinz (2010), Jardine (2016), and Jardine and Heinz (2016) did not aim to handle realistic corpora, but rather to investigate the necessary conditions for phonological learning. It would be uninsightful to abandon categorical approaches based on the performance of idealized learners designed for different research goals. Unlike previous proposals, the current study aims to bridge the gap between natural language data and theoretical studies based on categorical approaches such as (Jardine and Heinz, 2016).

The current paper focuses on the solution to the problem of exceptions and discusses the solution to accident gaps in §6. Exceptions can mislead previous categorical learning models that assume the equivalence between attestedness and grammaticality sequences. For example, a categorical learner assumes that attested sequences are all grammatical, it will incorrectly predict that the attested exception [sf] is grammatical in English. This problem in categorical approaches can be resolved by distinguishing exceptions from grammatical sequences based on the number of unique words in the lexicon that consists of the sequence, i.e. type frequency (Pierrehumbert, 2001; Albright and Hayes, 2003; Hayes and Londe, 2006; Hayes and Wilson, 2008; Albright, 2009). The primary assumption of this paper is that ungrammatical sound sequences, including exceptions, receive a lower lexical type frequency. In other words, ungrammatical sequences appear in fewer unique words than grammatical sequences. This *indirect negative evidence* of frequency information is crucial for human learners to extract constraints without being misled by exceptions (Regier and Gahl, 2004; Pearl and Lidz, 2009; Clark and Lappin, 2010; Pearl and Mis, 2016). Probabilistic learning models directly represent frequency/statistical information in grammar. In contrast, I propose a Neo-Trubetzkoyan approach that incorporates indirect negative evidence into categorical learning models. Essentially, this proposal iteratively identifies categorical constraints based on the comparison between Observed ($O$) and Expected ($E$) frequency and filters out ungrammatical sequences from the learning sample.

This proposal is named after Prince Nikolai Sergeyevich Trubetzkoy for his insights on the nature of frequency information (Trubetzkoy, 1939). Upon investigating the frequency of phoneme and phoneme combinations, Trubetzkoy observed that: "phoneme frequency is the result of a whole sequence of propelling forces." That is, besides the phonological grammar, the surface information of the phoneme frequency is also affected by extralinguistic factors. Therefore, "the absolute figures of the actual phoneme frequency are only of secondary importance. Only the *relationship* [emphasis added] of these figures to the theoretically expected figures of the frequency of the phoneme has real value" (Trubetzkoy, 1939, 1969). "The absolute figures of ac-

tual phoneme frequency" here means the observed frequency from the learning sample. In other words, surface frequency (attestedness) should not be directly interpreted as grammaticality, and the principled way of treating frequency information in phonological analysis is to focus on the relationship between $O$ and $E$. Trubetzkoy's observation is relevant to phonological acquisition because learners must also extract generalizations from attested forms.

## 2.4 Preliminaries of formal methods

This article only assumes basic knowledge of algebra (e.g. summation $\sum_{x\in\{1,\ldots,n\}} = x_1 + \ldots + x_n$) and set theory (element of $\in$; union $\cup$; intersection $\cap$; empty set {}; etc).

A language $L$ is a set of strings described by a grammar ($G$) (Heinz, 2011a,b). A Strictly 2-Local grammar $G$ ={*VV}, for example, characterizes a language that allows any strings except those with adjacent VV sequences. A string is a sequence of symbols $\sigma$ drawn from the alphabet (inventory) $\Sigma$ such as {C, V}. $\Sigma^*$ is the set of all possible strings for $\Sigma$. The content of the alphabet depends on the assumed phonological representations, such as segments and feature bundles. The current paper assumes a segmental representation and leaves the feature-based learning model to future work.

Given a string $s$, $\texttt{factor}(s, k)$ is a function that returns an *adjacent* sequence of length $k$, *a.k.a.* $k$-factors (Heinz, 2007, 2010). For example, $\texttt{factor}(\text{CCV}, 2)$ returns all 2-factors in CCV, including $\{\text{CC}, \text{CV}\}$. $k$-factors are constraints in (Tier-based) Strictly Local languages that capture most local and nonlocal phonotactic patterns and define a restrictive hypothesis space for efficient learning (Heinz, 2007; Heinz et al., 2011; Rogers and Pullum, 2011; Jardine and Heinz, 2016). A categorical Strictly $k$-Local grammar is similar to but different from the traditional $n$-gram model (Jelinek, 1998; Jurafsky and Martin, 2009), which predicts a probabilistic distribution. In contrast, a categorical Strictly $k$-Local grammar specifies a set of strings that ban a finite set of adjacent sequences of length $k$. Given $G$ ={*VV}, the string VVV, for example, would be ungrammatical, while CVC would be grammatical.

The current study assumes Strictly 2-Local Languages such as {*aa, *bb, ...} as the hypothesis space of phonotactic learning. This has been widely applied in previous works (Hayes and Wilson, 2008; Gouskova and Gallagher, 2020), and supported by typological studies based on *Subregular Hypothesis*: most phonological patterns reside in a restrictive subregular class of formal languages in Chomsky Hierarchy, especially Strictly 2-Local language (Rogers and Pullum, 2011; Heinz, 2018).

In a binary categorical grammar, the violation of the constraint $C$ for any string $s$ can be defined as follows: a string has one violation of the constraint if the constraint exists in the set of $k$-factors.

$$v(s, C) = \begin{cases} 0, & \text{if } \texttt{factor}(s, k) \cap \{C\} = \emptyset \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

Similarly, the grammatical score $g(s)$ for any string $s$ is defined as follows: the string is

grammatical ($g = 1$) if none of the constraints in $G$ exists in the set of $k$-factors.

$$g(s, G) = \begin{cases} 1, & \text{if } \texttt{factor}(s, k) \cap G = \emptyset \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $1$ and $0$ indicate grammatical and ungrammatical, respectively.

It is crucial to distinguish the categorical grammar introduced in this paper from a categorical OT grammar (Prince and Smolensky, 1993, *Optimality Theory*). Despite their similarities, they are different because the current paper ignores the ranking of phonotactic constraints and focuses on the acquisition of individual constraints. Moreover, constraint violations are nonculmulative in the current proposal because the author is not aware of evidence for the culmulativity effect in phonotactics.

Table 4 summarizes the notation and terminologies used in the paper.

| Notations | Terminologies | Examples |
|---|---|---|
| $\sigma$ | Symbol | §2; C, V |
| $\Sigma$ | Alphabet | §2; {C, V} |
| $\rtimes$ | Word boundary (left) | §6.2; |
| $\ltimes$ | Word boundary (right) | §6.2 |
| $\lvert x \rvert$ | The size of $x$ | §3; 2 for {C, V} |
| $s$ | a word/string from a finite sample | §3; CVC |
| $\Sigma^*$ | All possible strings for $\Sigma$ | §3 |
| $\Sigma^\ell$ | All possible strings of length $\ell$ for $\Sigma$ | §3 |
| $S$ | a finite sample of strings | §3; {CVC, CVV, VVC, VVV, VCV, CCV} |
| $G$ | the hypothesis grammar | §3; {*CC} |
| $\mathcal{T}$ | the target grammar | §3; {*CC} |
| $g$ | score of grammaticality | §2, §3; |
| $v$ | function of noncumulative constraint violations | §2, §3 |
| $w$ | constraint weight | §6.2 |
| $h$ | harmony score | §6.2 |
| $P^*$ | MaxEnt value | §6.2 |
| Pr | Probability | §3; §6.2 |
| $x \leftarrow y$ | Assigning value $y$ to $x$ | §8.1 |

Table 4: The list of essential notation

# 3 Proposal: Neo-Trubetzkoyan Phonotactic Learner

This section introduces the Neo-Trubetzkoyan Phonotactic learner. First, this section introduces the motivations underlying the computational learner; then, it explains the learning algorithm step by step through a toy example. The formal algorithm is defined in the Appendix.

## 3.1 The learning problem and toy example

How do human speakers learn to decide if a word is acceptable? This learning problem can be formalized as follows: how does a learner infer the target grammar $\mathcal{T}$ from the learning sample $\mathcal{S}$ which exemplifies the target language? A target grammar $\mathcal{T}$ is a set of constraints that describes the target language $\mathcal{L}$ and predicts whether a string is in the target language.

This paper uses the toy example below to illustrate the learning problem and learning models throughout the paper. Given the inventory {C, V}, the target grammar $\mathcal{T} = \{$*CC$\}$ underlies the target language $\mathcal{L}$ ={CV, VC, VV, CVC, ...} that penalizes the adjacent CC sequence. The toy learning sample $S$ consists of all three-length grammatical strings with one exception CCV which violates the target grammar. The word length is limited to three for the convenience of discussion, whereas the learner can handle samples with any combination of word lengths.

(1) Learning sample with one exception *CCV that violates the target grammar {*CC}:

$$S = \{\text{CVC, CVV, VVC, VVV, VCV, CCV}\}$$

The learner's input is the learning sample $S$, which only contains unique strings (types). $S$ includes grammatical strings extracted from the target language and a finite number of ungrammatical strings not in the target language. The learning sample only consists of positive evidence, which means that the learner does not have prior knowledge about lexical exceptions such as lexically specific indices (Moore-Cantwell and Pater, 2016).

After the learning procedure, the learner hypothesizes a categorical Strictly 2-Local grammar $G$ (hereafter "hypothesis grammar"), which consists of *inviolable* Strictly 2-Local constraints.

## 3.2 Overview of the Neo-Trubetzkoyan learner

In previous probabilistic learning models such as MaxEnt learner (Hayes and Wilson, 2008), the comparison between $O$ and $E$ is a *search heuristic* that selects parameters for weight optimization. This heuristic has been replaced by the *gain* criterion (Berger et al., 1996; Della Pietra et al., 1997) in recent MaxEnt learners (Berent et al., 2012; Wilson and Gallagher, 2018; Gouskova and Gallagher, 2020). $O < E$ indicates the restrictions on the sound sequence. Here, comparing $O$ and $E$ to select potential constraints is a key component, which provides a principled way to utilize indirect negative evidence from frequency information. This approach links the abstract representation in the grammar and speakers' knowledge of the frequency information in the surface forms.

The Neo-Trubetzkoyan learner iteratively filters out the ungrammatical attested exceptions based on the updated hypothesis grammar. The learning objective is to output a hypothesis grammar given the learning sample approximating the target grammar. Learning succeeds if the convergent grammar matches the target grammar that underlies the learning sample. Figure 3 shows the overview of the learning algorithm: After initializing a hypothesis grammar $G$ (Step 1), the learner compares the observed type frequency based on the learning sample and the expected type frequency based on the current hypothesis grammar (Step 2). Next, the learner stores potential constraints in a queue $Q$ based on the $O/E$ ratio (Step 3). Then, the learner updates

the hypothesis grammar with potential constraints (Step 4) and recalculates $O/E$ (return to Step 2). The loop continues until there are no more potential constraints ($Q = \emptyset$), and the learning converges.[1]

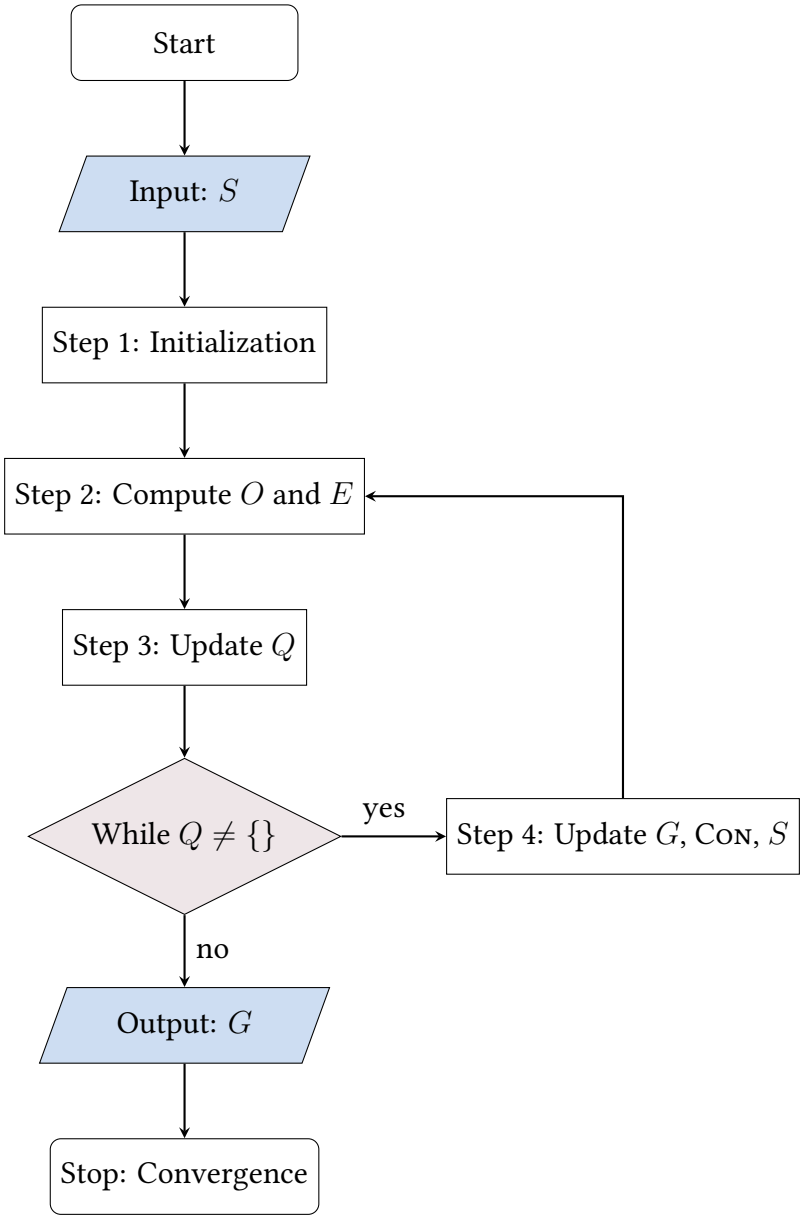



Figure 3: The learning procedure of the Neo-Trubetzkoyan learner

[1]The code demonstration can be accessed on the website: https://tinyurl.com/trubetzkoy.

### 3.3 Step 1: Initialization

Table 5 below shows the search tables that simulate phonotactic learning from the toy example. The left table shows the computations of $O$ and $E$, and the right table stores the variables: target grammar $\mathcal{T}$, hypothesis grammar $G$, hypothesis space Con, learning sample $S$, and queue $Q$ that stores potential constraints. The learner initializes all variables and an empty search table for the computation of $O$ and $E$. The initialized hypothesis grammar $G$ is an empty set, which means that the learner assumes that every possible sequence in $\Sigma^*$ is grammatical before learning begins. The hypothesis space Con includes all possible forbidden 2-factors.

| | $O$ | $E$ | $O-E$ |
|---|---|---|---|
| *VV | 0 | 0 | 0 |
| *VC | 0 | 0 | 0 |
| *CV | 0 | 0 | 0 |
| *CC | 0 | 0 | 0 |

| | | |
|---|---|---|
| $G$ | = | {} |
| Con | = | {*CV, *VV, *VC, *CC} |
| $S$ | = | {CVC, CVV, VVC, VVV, VCV, CCV} |
| $Q$ | = | {} |

Table 5: Initialization

### 3.4 Step 2 & 3 : Compute $O$ and $E$ and Update $Q$

After the initialization, the learner computes $O$ and $E$ for every constraint $C$ in Con based on the learning sample. I define the Observed (Type) Frequency ($O$) of a constraint $C$ as the number of unique strings in the sample that violates $C$:

$$O\left[C\right] = \sum_{s \in S} v(s, C) \tag{4}$$

In the toy sample, $S$ ={CVC, CVV, VVC, VVV, VCV, CCV}, $O$[*CC] $= 1$, $O$[*CV] $= 4$, $O$[*VC] $= 3$, $O$[*VV] $= 3$. Note that $O$[*VV] here is not 4 because the learner only counts unique words containing each 2-factor. Before computing $E$, if $O[C] = 0$ for any $C$, the learner adds $C$ to $G$, removes $C$ from Con, and assigns 0 to $E[C]$. I refer to this operation as REMOVE_GAPS (see the discussion on accidental gaps in §6).

#### 3.4.1 Approximating Expected (Type) Frequency

The Expected (Type) frequency $E[C]$ is the number of unique strings that violate a constraint $C$ in the hypothesis language $L$ described by the current hypothesis grammar $G$.[2] It is necessary to limit the length of strings in $L$; otherwise, the calculation of $E[C]$ becomes untractable because the hypothesis language could be infinite (Hayes and Wilson, 2008). In practice, the learner only counts the strings up to a maximum length $\ell_{\text{max}}$ from the longest string in the learning sample $S$. $S_\ell$ is a set of strings of length $\ell$ in $S$.

[2]This is different from the definition in Hayes and Wilson (2008), where the violations of constraints are cumulative, and $E[C]$ is the number of *violations* in the hypothesis language.

The learner approximates $E[C]$ as the sum of the Expected (Type) Frequency of strings of a given length $\ell$ in $\{1, \ldots, \ell_{\max}\}$:

$$E[C] \approx \sum_{l \in \{1,\ldots,\ell_{\max}\}} E_\ell[C] \tag{5}$$

First, the learner estimates the probability of strings that violate $C$ in $L$, then obtains $E_\ell[C]$ by multiplying this probability with $|S_\ell|$ the number of unique strings in $S_\ell$. The expected number of violations of constraint $C$ in $S_\ell$ is:

$$E_\ell[C] = |S_\ell| \times \mathsf{Pr}(C, G, \ell) \tag{6}$$

$\mathsf{Pr}(C, G, \ell)$ is the probability that a string violates $C$ in the expected sample $\Sigma^\ell$, which is computed with respect to Weighted Finite-state Acceptors (WFAs; Mohri, 2002; Riggle, 2004; Hayes and Wilson, 2008):

$$\mathsf{Pr}(C, G, \ell) = 1 - \frac{Z(\mathcal{N}'_\ell)}{Z(\mathcal{N}_\ell)}, \textit{ where } \mathcal{N}'_\ell \textit{ corresponds to } G' = G \cup \{C\} \tag{7}$$

WFA $\mathcal{N}_\ell$ represents all the possible strings with length $\ell$ that satisfy the current hypothesis grammar $G$. $Z(\mathcal{N}_\ell)$ counts strings that satisfy $G$ and receive zero penalty in $\mathcal{N}_\ell$. The WFA $\mathcal{N}'_\ell$ corresponds to the updated hypothesis grammar $G \cup \{C\}$, and $Z(\mathcal{N}'_\ell)$ is the count of allowed strings in the *updated* hypothesis grammar. The learner then calculates $Z$ with the algorithm for the Shortest-Distance Problems (Mohri, 2002; Gorman, 2016, `shortestdistance` in Pynini) that discovers the grammatical strings encoded in $\mathcal{N}_\ell$.

### 3.4.2 Exception-free examples

Table 6 lists (1) a hypothesis grammar $G$; (2) the idealized exception-free learning samples that match the expected samples for $G$; (3) $E$ values. For example, the exception-free learning sample [CCC, CCV, CVC, CVV, VVV, VCV, VCC, VVC] in the first row corresponds to an empty hypothesis grammar ($G = \{\}$). Recall Equation 6,

$$\begin{aligned} E_3[\text{*CC}] &= |S_3| \times \mathsf{Pr}(\text{*CC}, \{\}, 3) \\ &= |S_3| \times (1 - \frac{Z(\mathcal{N}'_3)}{Z(\mathcal{N}_3)}), \textit{ where } \mathcal{N}'_3 \textit{ corresponds to } G' = G \cup \{\text{*CC}\} \end{aligned} \tag{8}$$

$Z(\mathcal{N}_3) = 8$ because WFA $\mathcal{N}_3$ for $G = \{\}$ allows eight possible strings {CCC, CCV, VCC, CVC, CVV, VVV, VCV, VVC}. If *CC is added to the grammar, the updated grammar $G' = \{\text{*CC}\}$ only

allows five strings {CVC, CVV, VVV, VCV, VVC}. Therefore, the expected frequency of *CC is:

$$\begin{aligned} E_3\,[\text{*CC}] &= |S_3| \times (1 - \frac{Z(\mathcal{N}_3')}{Z(\mathcal{N}_3)}) \\ &= |S_3| \times (1 - \frac{5}{8}) \\ &= 8 \times (1 - \frac{5}{8}) \\ &= 8 \times \frac{3}{8} \\ &= 3 \end{aligned} \tag{9}$$

This correctly predicts the fact that three strings {CCC, CCV, VCC} violate the constraint *CC in the idealized learning sample.

| $G$ | Idealized learning sample $S_3$ | $E_3$[*CC] | $E_3$[*VV] | $E_3$[*CV] | $E_3$[*VC] |
|---|---|---|---|---|---|
| {} | {CCC, CCV, VCC, CVC, CVV, VVV, VCV, VVC} | 3 | 3 | 4 | 4 |
| {*CC} | {CVC, CVV, VVV, VCV, VVC} | 0 | 3 | 3 | 3 |
| {*VV} | {CCC, CCV, VCC, CVC, VCV} | 3 | 0 | 3 | 3 |
| {*CV} | {CCC, VCC, VVV, VVC} | 2 | 2 | 0 | 2 |
| {*VC} | {CCC, CCV, CVV, VVV} | 2 | 2 | 2 | 0 |
| {*CC, *VV} | {CVC, VCV} | 0 | 0 | 2 | 2 |
| {*CC, *VC} | {CVV, VVV} | 0 | 2 | 1 | 0 |
| {*CC, *CV} | {VVV, VVC} | 0 | 2 | 0 | 1 |
| {*CV, *VV} | {CCC, VCC} | 2 | 0 | 0 | 1 |
| {*CV, *VC} | {CCC, VVV} | 1 | 1 | 0 | 0 |
| {*VV, *VC} | {CCC, CCV} | 2 | 0 | 1 | 0 |
| {*CC, *VV, *CV} | {} | 0 | 0 | 0 | 0 |
| {*CC, *VV, *VC} | {} | 0 | 0 | 0 | 0 |
| {*CC, *VC, *CV} | {} | 0 | 0 | 0 | 0 |
| {*VV, *VC, *CV} | {} | 0 | 0 | 0 | 0 |
| {*CC, *VV, *CV, *VC} | {} | 0 | 0 | 0 | 0 |

Table 6: The list of idealized learning samples and corresponding hypothesis grammar, as well as expected frequencies for length 3

When one or more constraints are already present in the grammar, the learner removes any strings from the learning sample that violate the current hypothesis grammar $G$. For example, when $G = \{$*CC$\}$ (2nd row in Table 6), the idealized learning sample becomes {CVC, CVV, VVV, VCV, VVC} ($|S_3| = 5$. $E[$*CC$] = 0$ because *CC is already penalized by $G$ (*CC $\in G$). $E[$*VV$] = (1 - \frac{2}{5}) \cdot |S_3| = \frac{3}{5} \cdot 5 = 3$ because there are 5 strings allowed by $G = \{$*CC$\}$, and only 2 strings allowed by $G' = \{$*CC, *VV$\}$. Therefore, three of the strings allowed by {*CC} violate {*VV}.

### 3.4.3 Exceptionful examples

Consider the toy learning sample $S = \{$CVC, CVV, VVC, VVV, VCV, CCV$\}$ in Example 3.1, and let $G = \{\}$. First of all, $|S| = 6$. Secondly, for constraint *CC, $Z(\mathcal{N}) = 8$ and $Z(\mathcal{N}') = 5$ because

three strings violate the updated grammar $G' = \{CC\}$. The probability that a string violates *CC in the expected sample is the same as in the exception-free example above $\Pr(\text{*CC}, \{\}, 3) = 1 - \frac{5}{8} = \frac{3}{8}$. As a result, $E[\text{*CC}] = |S| \cdot \Pr(\text{*CC}, \{\}, 3) = 6 \cdot \frac{3}{8} = 2.25$. Table 7 shows the observed and expected frequency of all 2-factors in the toy example for an empty hypothesis grammar.

The learner finds the potential constraints where $O[C] < E[C]$ and the corresponding values $O - E$, which are temporarily stored in a queue $Q$. As illustrated in Table 7, for the toy sample, the learner discovers that $Q = \{(\text{*CC}, -1.25)\}$. This temporary "memorization" of the potential constraints is computationally bounded by the size of Con, which is bounded by $|\Sigma| \times k$ in an $\text{SL}_k$ grammar.

| | $O$ | $E$ | $O - E$ |
|---|---|---|---|
| *VV | 3 | 2.25 | 0.75 |
| *VC | 3 | 3 | 0 |
| *CV | 4 | 3 | 1 |
| *CC | 1 | 2.25 | −1.25 |

| | | |
|---|---|---|
| $G$ | = | {} |
| Con | = | {*CV, *VV, *VC, *CC} |
| $S$ | = | {CVC, CVV, VVC, VVV, VCV, CCV} |
| $Q$ | = | {(*CC, −1.25)} |

Table 7: Compute $O$ and $E$ and update $Q$

## 3.5 Step 4: Update $G$, Con, and $S$

While $Q \neq \{\}$, the learner updates $G$ by adding the constraint(s) with the smallest value $O - E$ $C_{\min(O-E)}$. The learner adds all constraints that share $\min(O - E)$ to $G$. Respectively, the learner removes $C_{\min(O-E)}$ from the hypothesis space Con. Moreover, the learner updates the learning sample $S$ by eliminating any string that violates the updated hypothesis grammar $G$.

As shown in Table 8, given $Q = \{(\text{*CC}, -1.25)\}$, the $\min(O - E) = -1.25 = O[\text{CC}] - E[\text{CC}]$, the learner adds *CC to $G$, removes *CC from Con, and eliminates CCV from $S$.

| | $O$ | $E$ | $O - E$ |
|---|---|---|---|
| *VV | 3 | 2.25 | 0.75 |
| *VC | 3 | 3 | 0 |
| *CV | 4 | 3 | 1 |
| *CC | 1 | 2.25 | −1.25 |

| | | |
|---|---|---|
| $G$ | = | {*CC} |
| Con | = | {*CV, *VV, *VC, ~~*CC~~} |
| $S$ | = | {CVC, CVV, VVC, VVV, VCV, ~~CCV~~} |
| $Q$ | = | {(*CC, −1.25)} |

Table 8: Update $G$, Con, and $S$

## 3.6 Iteration

After Step 4, the learner returns to Step 2 to recalculate $O$ and $E$. This compute-update-recompute iteration is crucial because $O$ and $E$ depend on the updated hypothesis grammar $G$. Iteration is repeated until no potential constraint is left, meaning that learning converges. As shown in Table 9, in the second iteration, after adding *CC to $G$, $O = E$ for all possible constraints, which means

$Q = \{\}$, the learned grammar successfully converges to {*CC} which matches the target grammar $\mathcal{T}$.

| | $O$ | $E$ | $O - E$ |
|---|---|---|---|
| *VV | 3 | 3 | 0 |
| *VC | 3 | 3 | 0 |
| *CV | 3 | 3 | 0 |

| | | |
|---|---|---|
| $G$ | = | {*CC} |
| Con | = | {*CV, *VV, *VC} |
| $S$ | = | {CVC, CVV, VVC, VVV, VCV} |
| $Q$ | = | {} |

Table 9: (2nd iteration) Compute $O$ and $E$ and update $Q$

### 3.7 Summary: Learning algorithm

In summary, the Neo-Trubetzkoyan learner starts by initializing an empty hypothesis grammar that allows for every possible sequence. As the learner accumulates indirect negative evidence from the learning sample, it gradually reduces the space of possible sequences and updates the hypothesis grammar $G$ with respect to the constraint selection criterion based on the comparison of the observed and expected type frequency. Assuming that the exceptions are of a lower observed type frequency than their expected frequency, the learner will penalize exceptions in the learned hypothesis grammar. The current proposal used well-studied algorithms based on WFAs to approximate the value $E$. I propose an iterative mechanism to incorporate the constraint selection criterion into the learner, which provides a principled way to utilize frequency information and handle exceptions in phonotactic learning.

## 4 Case study: Turkish

The following sections demonstrate the Neo-Trubetzkoyan learner in naturalistic corpora in Turkish and English. This section tests the learner's capability in capturing nonlocal vowel phonotactics from a naturalistic exceptionful Turkish corpus.

### 4.1 Target grammar: Turkish vowel phonotactic patterns

The current paper focuses on vowel phonotactic patterns in Turkish. Turkish vowels are shown in Table 10. Turkish orthography is converted to IPA, namely *ö* [ø], *ü* [y], *ı* [ɯ], and *a* [ɑ].

| | [−back] | | [+back] | |
|---|---|---|---|---|
| | [−round] | [+round] | [−round] | [+round] |
| [+high] | i | y | ɯ | u |
| [−high] | e | ø | ɑ | o |

Table 10: Turkish vowel system

In Turkish, two phonotactic constraints trigger progressive harmony across morpheme boundaries. First, a vowel cannot follow another vowel with a different value [back]. (a) and (b) in Table 11 show the [back] harmony in the Turkish nominative forms. For example, on *pullar* "girls", *lar* instead of *ler* surfaces due to the phonotactic constraint on the nonlocal *u...a* sequence.

However, this generalization has many exceptions, such as (c) and (d) in Table 11. Some exceptions are of very high token frequency in daily use, even higher than "grammatical" words that follow the vowel harmony pattern. For example, the exception *silah* "weapon" has a frequency of 26,658, while the grammatical word *sapık* 'pervert' is less frequent, with only $2,716$ occurrences in the Wiki corpus of ≈100 million words in subtitles.[3] This is why it is necessary to consider type instead of token frequency to identify exceptions (§3).

| | NOM.SG. | NOM.PL. | meaning | |
|---|---|---|---|---|
| a. | *ip* | *ipler* | "rope" | |
| | *köy* | *köyler* | "village" | (Clements et al., 1982) |
| | *yüz* | *yüzler* | "face" | |
| | *kız* | *kızlar* | "girl" | |
| | *pul* | *pullar* | "stamp" | |
| b. | *neden* | *nedenler* | "reason" | (Inkelas et al., 2000) |
| | *kiler* | *kilerler* | "pantry" | |
| | *pelür* | *pelürler* | "onionskin" | |
| | *boğaz* | *boğazlar* | "throat" | |
| | *sapık* | *sapıklar* | "pervert" | |
| c. | *mezar* | *mezarlar* | 'grave" | (Inkelas et al., 2000) |
| | *model* | *modeller* | "model" | |
| | *silah* | *silahlar* | "weapon" | |
| | *memur* | *memurlar* | "official" | |
| | *sabun* | *sabunlar* | "soap" | |
| d. | *etol* | *etoller* | "fur stole" | (Göksel and Kerslake, 2004) |
| | *saat* | *saatler* | "hour, clock" | |
| | *kahabat* | *kahabatler* | "fault" | |

Table 11: Turkish nominatives that undergo back harmony (a, b) and exceptions (c, d) (Gorman, 2013, P46)

Second, a high vowel cannot follow another vowel with a different [round] value, as shown in Table 12. However, the Turkish round harmony is well known for its exceptions, especially *labial attraction* where $a\mathrm{C}_{[+\text{labial}]}u$ is produced due to the intervocalic labial consonant, e.g. *sabur* "patient" (Lees, 1966). This pattern, however, is not internalized by native speakers (Zimmer, 1969). In other words, they are attested but ungrammatical (Gorman, 2013) exceptions to the round harmony patterns.

[3]Link: https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Turkish_WordList_10K

| | NOM.SG. | DAT.SG. | GEN.SG. | meaning | |
|---|---|---|---|---|---|
| a. | *ip* | *ipi* | *ipin* | “rope” | (Clements et al., 1982) |
| | *kız* | *kızı* | *kızın* | “girl” | |
| | *sap* | *sapı* | *sapın* | “stalk” | |
| | *köy* | *köyü* | *köyün* | “village” | |
| | *son* | *sonu* | *sonun* | “end” | |
| b. | *boğaz* | *boğazı* | *boğazın* | “throat” | (Inkelas et al., 2000) |
| | *pelür* | *pelürü* | *pelürün* | “onionskin” | |
| | *döviz* | *dövizi* | *dövizin* | “currency” | |
| | *yamuk* | *yamuğu* | *yamuğun* | “trapezoid” | |
| | *ümit* | *ümiti* | *ümitin* | “hope” | |

Table 12: Turkish nominal suffix allomorphy that undergoes round harmony (Gorman, 2013, P55)

Moreover, mid-round [-high, +round] vowels /ø/ and /o/ can only appear in the initial positions, as in *ödev* “homework” and *oyun* “game”. This means that they should not follow any vowels, e.g. *a...ø, *e...o, etc. In summary:

1. A vowel cannot follow another vowel with a different [back] value;
2. A high vowel cannot follow another vowel with a different [round] value.
3. Mid-round vowels cannot follow any vowels.

These generalizations constitute the *target grammar* for phonotactic learning, as illustrated in Table 13. Learning is successful if the learned grammar maximally approximates the target grammar.

| $\sigma_1 \downarrow \sigma_2 \rightarrow$ | i | e | y | ø | ɯ | a | u | o |
|---|---|---|---|---|---|---|---|---|
| i | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| e | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| y | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| ø | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| ɯ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| u | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| o | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

Table 13: Target grammar of Turkish vowel phonotactics; cells for grammatical 2-factors are highlighted

## 4.2 Turkish learning result

I applied the proposed algorithm to the Turkish Electronic Living Lexicon (TELL; Inkelas et al., 2000; Gouskova and Stanton, 2021), which consists of $\approx$ 66000 roots and the elicited derived forms. Table 14 shows the raw frequency of all 2-factors in TELL. Grammatical vowel phonotactics in previous literature are highlighted. The occurrences of ungrammatical 2-factors (nonzero frequency) are the exceptions of the phonotactic patterns.

| $\sigma_1 \downarrow \sigma_2 \rightarrow$ | i | e | y | ø | ɯ | a | u | o |
|---|---|---|---|---|---|---|---|---|
| i | 10950 | 4768 | 221 | 123 | 768 | 3216 | 202 | 1000 |
| e | 15984 | 7130 | 591 | 129 | 663 | 2873 | 625 | 760 |
| y | 422 | 2944 | 2465 | 43 | 121 | 750 | 177 | 59 |
| ø | 32 | 982 | 1179 | 27 | 19 | 98 | 18 | 19 |
| ɯ | 247 | 392 | 17 | 60 | 6360 | 3009 | 93 | 207 |
| a | 4369 | 3197 | 394 | 308 | 16887 | 10267 | 1526 | 1656 |
| u | 475 | 606 | 147 | 40 | 153 | 3035 | 4058 | 155 |
| o | 857 | 787 | 139 | 42 | 99 | 2591 | 3737 | 684 |

Table 14: The raw frequency of 2-factors in the learning sample; cells of documented grammatical 2-factors are highlighted.

Table 15 shows the comparison between the learned grammar in the Neo-Trubetzkoyan learner (a) and the target grammar (b). The learned grammar approximates the target grammar except for several mismatches, in which the raw frequency is misleading. In the future, it is worth to verify in experiments whether these mismatches come from loanwords or compounds that do not undergo the vowel harmony pattern.

| $\sigma_1 \downarrow \sigma_2 \rightarrow$ | i | e | y | ø | ɯ | a | u | o |
|---|---|---|---|---|---|---|---|---|
| i | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| e | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| y | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ø | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ɯ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| a | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| u | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| o | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

(a) Neo-Trubetzkoyan

| $\sigma_1 \downarrow \sigma_2 \rightarrow$ | i | e | y | ø | ɯ | a | u | o |
|---|---|---|---|---|---|---|---|---|
| i | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| e | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| y | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| ø | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| ɯ | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| a | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| u | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| o | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

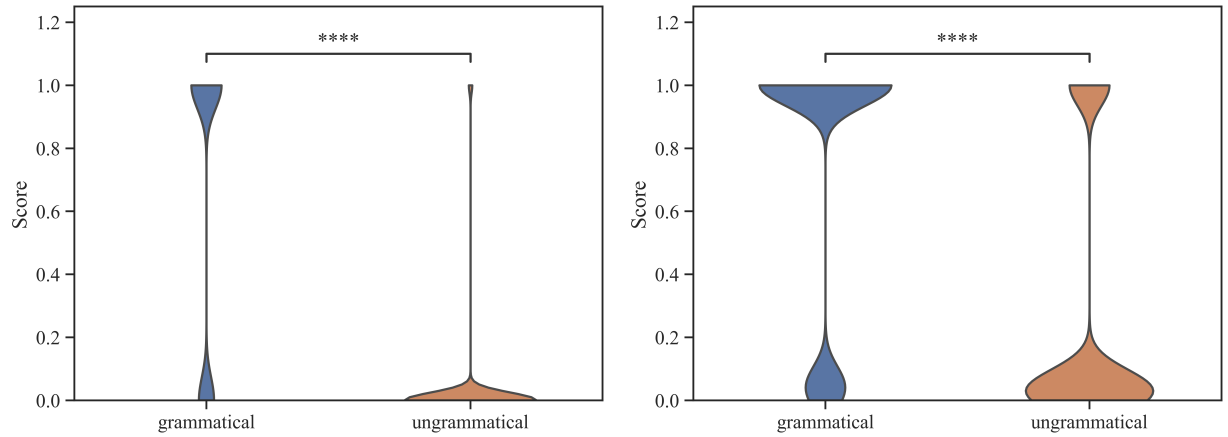
(b) Target grammar

Table 15: Learned grammar (a) vs. target grammar (b)

## 4.3 Model evaluation: predicting categorical acceptability

To compare with Hayes and Wilson (2008)'s MaxEnt learner, the current proposal and MaxEnt learner are both trained on TELL and evaluated on $20,000$ manually labeled nonce words in Turkish. I created the test data for the Turkish categorical judgment of nonce words based on the generalization from the previous literature (Lees, 1966; Zimmer, 1969; Gorman, 2013) because human judgments are unavailable. This method was employed in modeling to evaluate the learner's performance on low-resource languages such as Quechua when experimental data is absent (Gouskova and Gallagher, 2020; Gallagher et al., 2019).

I map the categorical judgments to scores 0 and 1. Figure 4 shows the clustering of predicted grammatical score groups in both the Neo-Trubetzkoyan (a) and the MaxEnt (b) learner, grouped by grammatical labels. I applied a nonparametric Mann-Whitney-Wilcoxon test to the pairs of "grammatical" vs. "ungrammatical" words, and both learners significantly distinguish the distributions of these pairs ($p < 0.001$). This means that the predicted scores in both learners can predict the categorical generalization reflected in previous literature (Lees, 1966; Zimmer, 1969).




(a) Neo-Trubetzkoyan

(b) MaxEnt

Figure 4: The clustering of the predicted scores of Turkish nonce words by the Neo-Trubetzkoyan (a) and MaxEnt learner (b) , grouped by the label of grammaticality

Furthermore, I computed the $F$-score and the binary precision, which measure the performance of the models in the binary classification, as illustrated in Table 16. The Neo-Trubetzkoyan learner achieved higher scores in all measures except the slightly lower binary accuracy of grammatical words due to the generalization errors described above.

| | | Neo-Trubetzkoyan | MaxEnt |
|---|---|---|---|
| $F$-score | | **0.790** | 0.667 |
| binary accuracy | grammatical | 0.666 | **0.761** |
| | ungrammatical | **0.974** | 0.770 |

Table 16: $F$-score and binary accuracy for binary classification; higher values are highlighted

I also measured the correlation score between the grammaticality labeled (0/1) and the judgments predicted by both learners. As illustrated in Table 17, the Neo-Trubetzkoyan learner is superior to the MaxEnt learner in all measures.

| | Neo-Trubetzkoyan | MaxEnt |
|---|---|---|
| Pearson correlation | **0.673** | 0.535 |
| Spearman correlation | **0.673** | 0.565 |
| Kendall correlation | **0.673** | 0.510 |

Table 17: Correlation scores of the predicted judgment and the labels in vowel nonce words; higher values are highlighted

To summarize, I trained the Neo-Trubetzkoyan learner using an exceptionful large-scale corpus of Turkish forms. The learner not only significantly distinguished the distributions of grammatical and ungrammatical words in Turkish nonce words but also achieved a high correlation between the predicted judgment and the labels of grammaticality supported by previous literature (Lees, 1966; Zimmer, 1969; Gorman, 2013). The Neo-Trubetzkoyan learner is superior to the MaxEnt learner (Hayes and Wilson, 2008) in capturing Turkish phonotactic patterns in the presence of exceptions.

# 5 Case study: English onsets

the proposed learner is further applied to the learning sample and behavioral data of English onsets. The learning result successfully predicts the gradient well-formedness judgments of English nonce words indicated by Likert ratings (Daland et al., 2011).

## 5.1 English learning sample

The current proposal can be applied to both exceptionful and exception-free corpora. I first trained the learner on the exception-free learning sample in which exceptions such as [sf] (as in *sphere*) and [pw] (*Puerto Rico*) are removed. This corpus consists of 31641 word-initial onsets from word types (instead of tokens) in the online CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/), as illustrated in Table 18:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| k | 2764 | p ɹ | 1046 | tʃ | 329 | d ɹ | 211 | θ ɹ | 73 |
| ɹ | 2752 | w | 780 | b ɹ | 319 | k w | 201 | s k w | 69 |
| d | 2526 | n | 716 | s p | 313 | s t ɹ | 183 | t w | 55 |
| s | 2215 | v | 615 | f l | 290 | θ | 173 | s p ɹ | 51 |
| m | 1965 | g | 537 | k l | 285 | s w | 153 | ʃ ɹ | 40 |
| p | 1881 | dʒ | 524 | s k | 278 | g l | 131 | s p l | 27 |
| b | 1544 | s t | 521 | j | 268 | h w | 111 | ð | 19 |
| l | 1225 | t ɹ | 515 | f ɹ | 254 | s n | 109 | d w | 17 |
| f | 1222 | k ɹ | 387 | p l | 238 | s k ɹ | 93 | g w | 11 |
| h | 1153 | ʃ | 379 | b l | 213 | z | 83 | θ w | 4 |
| t | 1146 | g ɹ | 331 | s l | 213 | s m | 82 | s k l | 1 |

Table 18: Frequency of English onsets in the exception-free learning sample

I trained the learner on the exceptionful learning sample, which consists of onsets from 31985 unique words in the CMU Pronouncing Dictionary that have at least one occurrence in the CELEX database (Hayes and Wilson, 2008).[4] A majority of this dataset overlaps with the exception-free learning sample shown in Table 18. Table 19 shows the frequency of additional onsets in the exceptionful learning sample.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| f j | 55 | s f | 5 | ʃ w | 3 | n w | 1 |
| m j | 54 | s p j | 5 | ʒ | 3 | p w | 1 |
| h j | 50 | ʃ l | 5 | f w | 2 | s ɹ | 1 |
| k j | 45 | ʃ m | 5 | g j | 2 | s θ | 1 |
| p j | 34 | n j | 4 | k n | 2 | ʃ p | 1 |
| b j | 21 | s k j | 4 | v l | 2 | v ɹ | 1 |
| d j | 9 | ʃ n | 4 | z j | 2 | z l | 1 |
| t j | 6 | b w | 3 | h ɹ | 1 | z w | 1 |
| v j | 6 | ʃ t | 3 | m w | 1 | | |

Table 19: Frequency of additional English onsets in the exceptionful learning sample

## 5.2 English learning result

The learner initializes $22 * 22 = 444$ 2-factors for 22 consonants that occur at the beginning, excluding [x] (as in *loch*) and [ŋ] (*ring*). The learner was run five times. The learning results are consistent because the constraint selection involves minimum stochastic operations ($Z$ in $E$ approximation), in contrast to learners based on stochastic gradient descent (Hayes and Wilson, 2008).

Table 20 shows the learned grammar $G$ after learning from exception-free learning sample. The column to the left indicates the first symbol in a 2-factor, and the top row indicates the

[4]Special thanks to Bruce Hayes, who provided this training data.

second symbol. Grammatical 2-factors (highlighted cells) are indicated by 1, and ungrammatical ones are indicated by 0. For example, [mm] is predicted as ungrammatical and receives 0, while [kl] is predicted as grammatical and receives 1.

| $\sigma_1 \downarrow \sigma_2 \rightarrow$ | p | b | t | d | k | g | f | v | s | z | θ | ð | ʃ | tʃ | dʒ | j | ɹ | l | m | n | w | h |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| t | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| k | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| g | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| f | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| v | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| z | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| θ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| ð | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ʃ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| tʃ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| dʒ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| j | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ɹ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| l | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| m | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| n | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| w | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| h | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Table 20: The learned grammar $G$ for English onsets (exception-free learning sample). Highlighted cells indicate grammatical 2-factors ($g(s, G) = 1$).

Although attestedness cannot imply grammaticality, the grammatical factors 2 predicted by the learned grammar here are all attested, e.g. [kl] (*freq.* = 285), [bl] (*freq.* = 213), and [sw] (*freq.* = 153). Among the penalized 2-factors, only [dw], [gw], and [θw] are *exceptions*, which are attested in the learning data but predicted as ungrammatical by the learner.

Table 21 shows the learned grammar induced from the exceptionful learning sample. Compared to Table 20, the Neo-Trubetzkoyan learner discovers several additional constraints, including *sn, *sw (*freq.*= 153), *hw (*freq.*= 111), *ʃɹ (*freq.*= 40), *θɹ (*freq.*= 73), *tw (*freq.*= 55) and *kw (*freq.*= 201). It seems that the learner tends to restrict the grammar even more when it is exposed to exceptions. Moreover, the learned grammar successfully eliminates 2-factors such as *fj in the exceptions with low type frequency in CELEX dataset, as illustrated in Table 19.

| $\sigma_1 \downarrow \sigma_2 \rightarrow$ | p | b | t | d | k | g | f | v | s | z | ʃ | ʒ | θ | ð | tʃ | dʒ | j | ɹ | l | m | n | w | h |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| p | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| t | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| k | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| g | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| f | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| v | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| s | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| z | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| θ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ð | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ʃ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ʒ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| tʃ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| dʒ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| j | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ɹ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| l | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| m | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| n | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| w | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| h | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 21: The learned grammar $G$ for English onsets (exceptionful learning sample). Highlighted cells indicate grammatical 2-factors ($g(s, G) = 1$).

## 5.3 Model evaluation: predicting gradient acceptability

This section evaluated the learned grammar in the test data with 96 nonce words of the CCVCVC structure, in which a word-initial onset CC is concatenated with a *tail* VCVC (Daland et al., 2011). They used 48 distinct onsets paired with six distinct tails, namely *eegiff*. *ezzig*, *eppid*, *ottiff*, *eebid*, and *ossip*.

This test data is one of the few available datasets with gradient acceptability of nonce words from human behavioral experiments. Each nonce word was rated on a Likert scale from 1 (unlikely) to 6 (likely) by participants with high English proficiency recruited from the Mechanical Turk (Daland et al., 2011). Daland et al. (2011) manually labeled the attestedness (attested $\gg$ marginal $\gg$ unattested) of the onsets.

### 5.3.1 Exception-free learning sample

Table 23 shows the onsets, orthographies presented to the participants, the attestedness, average Likert ratings, and the predicted judgment/grammaticality ($g$) by the proposed learner.

The learned grammar assigns ungrammatical to most marginal forms $g = 0$, except [tw] (#87, #88), which has a relatively high frequency in the learning sample. Another mismatch between the predicted judgment and attestedness is in the attested [θw] (#81, #82). As shown in Table 22, [θw] is considered attested in Daland et al. (2011) but extremely rare (*freq.*= 4) in the learning sample.

| onset | attestedness | $g$ | frequency | Likert |
|---|---|---|---|---|
| tw | marginal | 1 | 55 | 3.4 or 3.5 |
| θw | attested | 0 | 4 | 2.2 or 2.65 |

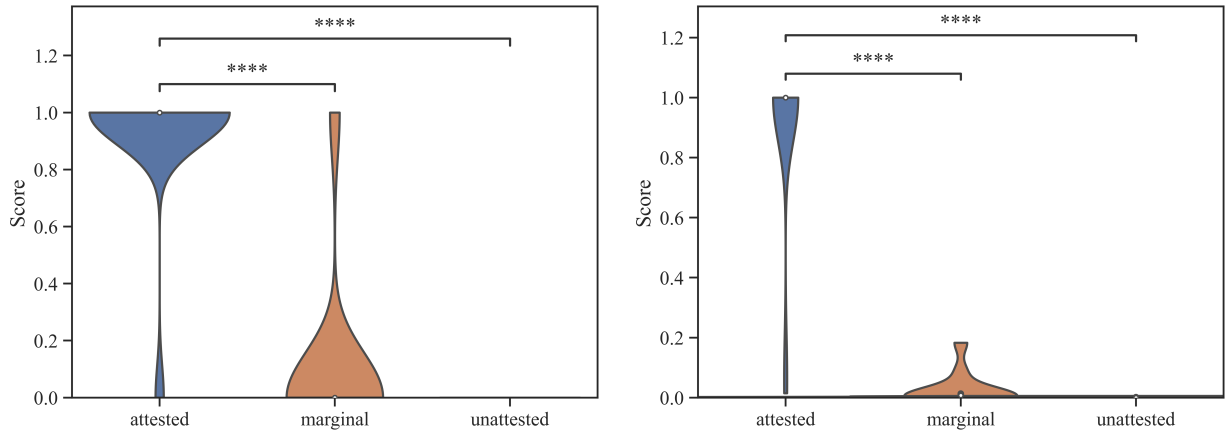
Table 22: Analysis of the mismatches between predicted grammaticality and attestedness in Daland et al. (2011)

| # | onset | orthography | attestedness | Likert | $g$ | # | onset | orthography | attestedness | Likert | $g$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | bl | *bleegiff* | attested | 3.8 | 1 | 49 | pk | *pkeebid* | unattested | 1.65 | 0 |
| 2 | bl | *blezzig* | attested | 3.35 | 1 | 50 | pk | *pkossip* | unattested | 1.55 | 0 |
| 3 | bɹ | *breegiff* | attested | 3.6 | 1 | 51 | pl | *pleppid* | attested | 3.85 | 1 |
| 4 | bɹ | *breppid* | attested | 3.9 | 1 | 52 | pl | *plossip* | attested | 4.35 | 1 |
| 5 | bw | *bweegiff* | marginal | 2.25 | 0 | 53 | pɹ | *prezzig* | attested | 3.65 | 1 |
| 6 | bw | *bwossip* | marginal | 2.7 | 0 | 54 | pɹ | *prottiff* | attested | 4.25 | 1 |
| 7 | dg | *dgeppid* | unattested | 1.42 | 0 | 55 | pw | *pweegiff* | unattested | 2.3 | 0 |
| 8 | dg | *dgottiff* | unattested | 1.45 | 0 | 56 | pw | *pwezzig* | unattested | 2.15 | 0 |
| 9 | dn | *dneegiff* | unattested | 1.63 | 0 | 57 | rd | *rdeegiff* | unattested | 1.15 | 0 |
| 10 | dn | *dnottiff* | unattested | 1.6 | 0 | 58 | rd | *rdossip* | unattested | 1.5 | 0 |
| 11 | dɹ | *dreegiff* | attested | 3.15 | 1 | 59 | rg | *rgeebid* | unattested | 1.6 | 0 |
| 12 | dɹ | *dreppid* | attested | 4.35 | 1 | 60 | rg | *rgeppid* | unattested | 1.45 | 0 |
| 13 | dw | *dwezzig* | marginal | 2.8 | 0 | 61 | rl | *rleegiff* | unattested | 1.8 | 0 |
| 14 | dw | *dwottiff* | marginal | 2.3 | 0 | 62 | rl | *rlezzig* | unattested | 1.45 | 0 |
| 15 | fl | *flezzig* | attested | 3.95 | 1 | 63 | rn | *rneppid* | unattested | 1.65 | 0 |
| 16 | fl | *flottiff* | attested | 4.25 | 1 | 64 | rn | *rnossip* | unattested | 1.5 | 0 |
| 17 | fn | *fneebid* | unattested | 1.7 | 0 | 65 | ʃl | *shleebid* | marginal | 3.1 | 0 |
| 18 | fn | *fnezzig* | unattested | 1.7 | 0 | 66 | ʃl | *shlezzig* | marginal | 3.15 | 0 |
| 19 | fɹ | *freppid* | attested | 4.35 | 1 | 67 | ʃm | *shmeegiff* | marginal | 2.7 | 0 |
| 20 | fɹ | *frossip* | attested | 4.7 | 1 | 68 | ʃm | *shmottiff* | marginal | 2.65 | 0 |
| 21 | fw | *fweebid* | marginal | 2.55 | 0 | 69 | ʃn | *shneegiff* | marginal | 2.4 | 0 |
| 22 | fw | *fwezzig* | marginal | 2.25 | 0 | 70 | ʃn | *shneppid* | marginal | 2.79 | 0 |
| 23 | gl | *gleppid* | attested | 3.9 | 1 | 71 | ʃɹ | *shreebid* | attested | 4.05 | 1 |
| 24 | gl | *glottiff* | attested | 3.55 | 1 | 72 | ʃɹ | *shreppid* | attested | 4 | 1 |
| 25 | gɹ | *greebid* | attested | 4.85 | 1 | 73 | ʃw | *shweppid* | marginal | 2.9 | 0 |
| 26 | gɹ | *grottiff* | attested | 4.15 | 1 | 74 | ʃw | *shwossip* | marginal | 3 | 0 |
| 27 | gw | *gweebid* | marginal | 2.85 | 0 | 75 | sm | *smeebid* | attested | 4.05 | 1 |
| 28 | gw | *gwottiff* | marginal | 2.5 | 0 | 76 | sm | *smottiff* | attested | 3.8 | 1 |
| 29 | kl | *kleebid* | attested | 3.95 | 1 | 77 | sn | *sneegiff* | attested | 3.5 | 1 |
| 30 | kl | *klossip* | attested | 4.05 | 1 | 78 | sn | *snossip* | attested | 4.45 | 1 |
| 31 | km | *kmeppid* | unattested | 1.3 | 0 | 79 | sw | *sweegiff* | attested | 3 | 1 |
| 32 | km | *kmossip* | unattested | 1.85 | 0 | 80 | sw | *swezzig* | attested | 3.4 | 1 |
| 33 | kɹ | *kreebid* | attested | 4.25 | 1 | 81 | θw | *thweppid* | attested | 2.2 | 0 |
| 34 | kɹ | *krezzig* | attested | 3.3 | 1 | 82 | θw | *thwossip* | attested | 2.65 | 0 |
| 35 | kw | *kweebid* | attested | 3.5 | 1 | 83 | tl | *tleebid* | unattested | 1.75 | 0 |
| 36 | kw | *kwottiff* | attested | 2.5 | 1 | 84 | tl | *tlottiff* | unattested | 1.84 | 0 |
| 37 | lm | *lmeebid* | unattested | 1.5 | 0 | 85 | tɹ | *trezzig* | attested | 4.05 | 1 |
| 38 | lm | *lmottiff* | unattested | 1.3 | 0 | 86 | tɹ | *trossip* | attested | 5 | 1 |
| 39 | ln | *lneegiff* | unattested | 1.45 | 0 | 87 | tw | *tweegiff* | marginal | 3.4 | 1 |
| 40 | ln | *lnezzig* | unattested | 1.45 | 0 | 88 | tw | *twossip* | marginal | 3.5 | 1 |
| 41 | lt | *ltezzig* | unattested | 1.55 | 0 | 89 | vl | *vleppid* | marginal | 2.3 | 0 |
| 42 | lt | *ltottiff* | unattested | 1.4 | 0 | 90 | vl | *vlossip* | marginal | 3.7 | 0 |
| 43 | ml | *mleppid* | unattested | 1.5 | 0 | 91 | vɹ | *vreebid* | marginal | 2.7 | 0 |
| 44 | ml | *mlossip* | unattested | 1.8 | 0 | 92 | vɹ | *vrezzig* | marginal | 2.3 | 0 |
| 45 | mɹ | *mreegiff* | unattested | 2 | 0 | 93 | vw | *vweegiff* | unattested | 1.7 | 0 |
| 46 | mɹ | *mrottiff* | unattested | 1.7 | 0 | 94 | vw | *vwottiff* | unattested | 1.55 | 0 |
| 47 | nl | *nleebid* | unattested | 1.85 | 0 | 95 | zɹ | *zreppid* | unattested | 2.05 | 0 |
| 48 | nl | *nlezzig* | unattested | 1.35 | 0 | 96 | zɹ | *zrossip* | unattested | 2.1 | 0 |

Table 23: The predicted judgment and corresponding data in English nonce words; learned from the exception-free learning sample.

I trained the MaxEnt learner with the standard settings for comparison. I set the parameter "Maximum $O/E$" to $1$ and $0.01$, and achieved similar results. Only the best results are reported here. I avoid using the *temperature* parameter to fine-tune the predictions (Hayes and Wilson, 2008). Only the predicted MaxEnt values $P^*$ are reported below because the normalized word probabilities are too small.

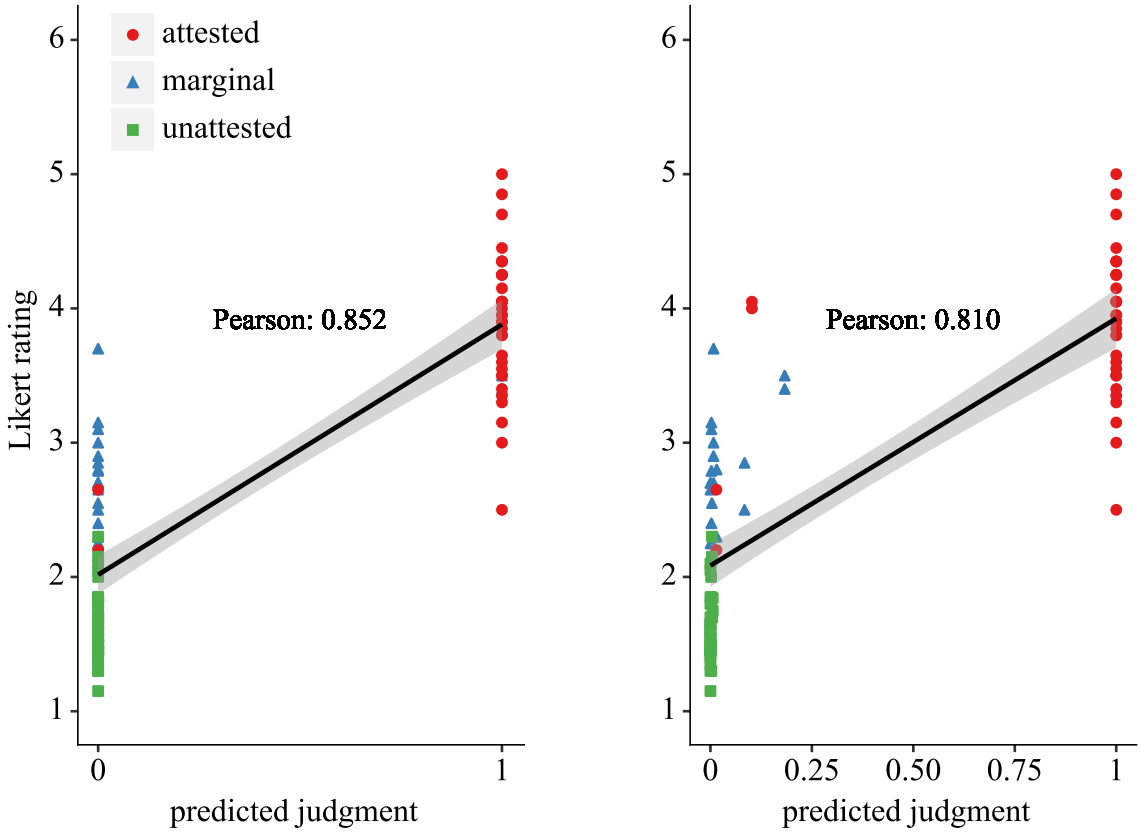
Figure 5 shows the clustering of predicted grammaticality scores in both Neo-Trubetzkoyan (a) and MaxEnt (b) learners, grouped by category of attestedness. I applied a nonparametric Mann-Whitney-Wilcoxon test to the pairs of "attested" vs. "marginal" and "attested" vs. 'unattested' words. Both learners significantly distinguish the distributions of these pairs ($p < 0.001$). As mentioned above, these labels of attestedness are validated by the participants' judgment. This means that the predicted scores in both learners can predict the categorical attestedness of onsets that reflects the participants' categorical judgments.




(a) Neo-Trubetzkoyan (b) MaxEnt ($P^*$)

Figure 5: The clustering of the predicted scores of English nonce words by the Neo-Trubetzkoyan (a) and MaxEnt learner (b), grouped by the label of attestedness.

Figure 6 illustrates the correlation between human judgment and the predicted scores of grammaticality by the Neo-Trubetzkoyan (a) and MaxEnt learner (b; UCLA Phonotactic learner; Hayes and Wilson, 2008). The Neo-Trubetzkoyan learner ($0$ for ungrammatical, $1$ for grammatical) achieved $0.852$ in Pearson's $r$ test, which is higher than $0.810$ in the MaxEnt learner result.




(a) Neo-Trubetzkoyan

(b) MaxEnt ($P^*$)

Figure 6: The correlation between the Likert rating and the predicted judgments by the Neo-Trubetzkoyan (a) and MaxEnt learner (b)

the proposed learner also achieved high scores on the Spearman and Kendall correlation test, close to the MaxEnt learner, as illustrated in Table 24. The $p$ values for all tests are $< 0.001$ (highly significant correlation).

| | Neo-Trubetzkoyan | MaxEnt |
|---|---|---|
| Pearson correlation | **0.852** | 0.810 |
| Spearman correlation | 0.816 | **0.883** |
| Kendall correlation | 0.674 | **0.707** |

Table 24: Correlation scores of predicted judgment (from the exception-free learning sample) and human judgment (Likert ratings) in English nonce words

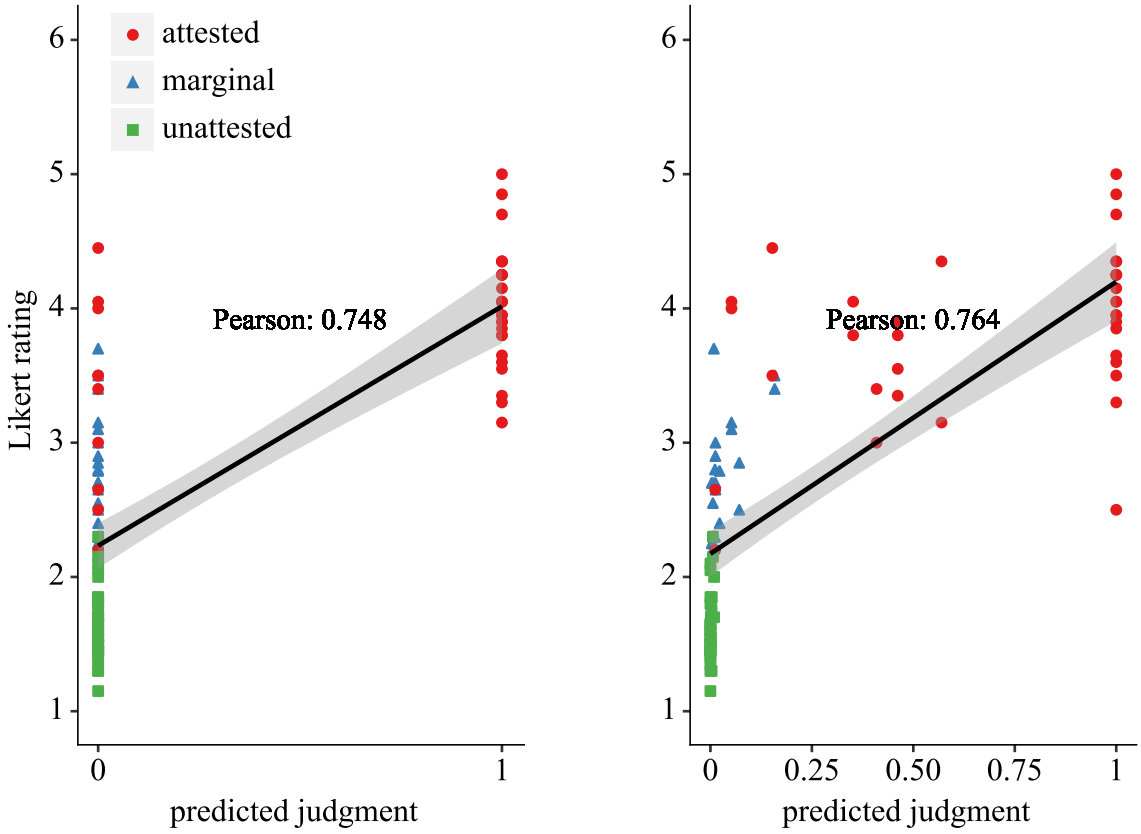
### 5.3.2 Exceptionful learning sample

It is not ideal to evaluate the learning result of the exceptionful learning sample on the behavioral data reported in Daland et al. (2011). This is because Daland et al. (2011) did not include a large number of exceptions reported in the CELEX data set (Table 19). In the future, it is worth collecting new behavioral data that specifically target phonotactic exceptions. Despite this issue, the current article will still report the evaluation result based on the gradient acceptability in Daland et al. (2011) below as a reference.

The correlation scores of two learners are illustrated in Table 25. The $p$ values for all tests are $< 0.001$ (highly significant correlation). As shown in Figure 7, both the Neo-Trubetzkoyan ($\rho = 0.748$) and MaxEnt ($\rho = 0.764$) learners achieved high correlation scores between the predicted judgments and the Likert ratings.

| | Neo-Trubetzkoyan | MaxEnt |
|---|---|---|
| Pearson correlation | 0.748 | **0.764** |
| Spearman correlation | 0.709 | **0.914** |
| Kendall correlation | 0.586 | **0.763** |

Table 25: Correlation scores of predicted judgment (from the exceptionful learning sample) and Likert ratings in English nonce words




(a) Neo-Trubetzkoyan

(b) MaxEnt ($P^*$)

Figure 7: The correlation between the Likert ratings and the predicted judgments by the Neo-Trubetzkoyan (a) and MaxEnt learner (b); learned from the exceptionful learning sample

Although the MaxEnt learner seems superior in all correlation scores, the mistakes made by the Neo-Trubetzkoyan learner are informative. As illustrated in Table 26, the Neo-Trubetzkoyan learner made several new predictions that affect their overall performance. For example, well-attested sequences *kw (*freq* $= 201$), *sn (*freq* $= 109$), and *sw (*freq* $= 153$) are misgeneralized as ungrammatical ($g = 0$). These errors are likely due to the rigorous $O < E$ criterion, which should be improved in the future.

| # | onset | orthography | attestedness | Likert | $g$ | # | onset | orthography | attestedness | Likert | $g$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | bl | *bleegiff* | attested | 3.8 | 1 | 49 | pk | *pkeebid* | unattested | 1.65 | 0 |
| 2 | bl | *blezzig* | attested | 3.35 | 1 | 50 | pk | *pkossip* | unattested | 1.55 | 0 |
| 3 | bɹ | *breegiff* | attested | 3.6 | 1 | 51 | pl | *pleppid* | attested | 3.85 | 1 |
| 4 | bɹ | *breppid* | attested | 3.9 | 1 | 52 | pl | *plossip* | attested | 4.35 | 1 |
| 5 | bw | *bweegiff* | marginal | 2.25 | 0 | 53 | pɹ | *prezzig* | attested | 3.65 | 1 |
| 6 | bw | *bwossip* | marginal | 2.7 | 0 | 54 | pɹ | *prottiff* | attested | 4.25 | 1 |
| 7 | dg | *dgeppid* | unattested | 1.42 | 0 | 55 | pw | *pweegiff* | unattested | 2.3 | 0 |
| 8 | dg | *dgottiff* | unattested | 1.45 | 0 | 56 | pw | *pwezzig* | unattested | 2.15 | 0 |
| 9 | dn | *dneegiff* | unattested | 1.63 | 0 | 57 | rd | *rdeegiff* | unattested | 1.15 | 0 |
| 10 | dn | *dnottiff* | unattested | 1.6 | 0 | 58 | rd | *rdossip* | unattested | 1.5 | 0 |
| 11 | dɹ | *dreegiff* | attested | 3.15 | 1 | 59 | rg | *rgeebid* | unattested | 1.6 | 0 |
| 12 | dɹ | *dreppid* | attested | 4.35 | 1 | 60 | rg | *rgeppid* | unattested | 1.45 | 0 |
| 13 | dw | *dwezzig* | marginal | 2.8 | 0 | 61 | rl | *rleegiff* | unattested | 1.8 | 0 |
| 14 | dw | *dwottiff* | marginal | 2.3 | 0 | 62 | rl | *rlezzig* | unattested | 1.45 | 0 |
| 15 | fl | *flezzig* | attested | 3.95 | 1 | 63 | rn | *rneppid* | unattested | 1.65 | 0 |
| 16 | fl | *flottiff* | attested | 4.25 | 1 | 64 | rn | *rnossip* | unattested | 1.5 | 0 |
| 17 | fn | *fneebid* | unattested | 1.7 | 0 | 65 | ʃl | *shleebid* | marginal | 3.1 | 0 |
| 18 | fn | *fnezzig* | unattested | 1.7 | 0 | 66 | ʃl | *shlezzig* | marginal | 3.15 | 0 |
| 19 | fɹ | *freppid* | attested | 4.35 | 1 | 67 | ʃm | *shmeegiff* | marginal | 2.7 | 0 |
| 20 | fɹ | *frossip* | attested | 4.7 | 1 | 68 | ʃm | *shmottiff* | marginal | 2.65 | 0 |
| 21 | fw | *fweebid* | marginal | 2.55 | 0 | 69 | ʃn | *shneegiff* | marginal | 2.4 | 0 |
| 22 | fw | *fwezzig* | marginal | 2.25 | 0 | 70 | ʃn | *shneppid* | marginal | 2.79 | 0 |
| 23 | gl | *gleppid* | attested | 3.9 | 1 | 71 | ʃɹ | *shreebid* | attested | 4.05 | 1 |
| 24 | gl | *glottiff* | attested | 3.55 | 1 | 72 | ʃɹ | *shreppid* | attested | 4 | 1 |
| 25 | gɹ | *greebid* | attested | 4.85 | 1 | 73 | ʃw | *shweppid* | marginal | 2.9 | 0 |
| 26 | gɹ | *grottiff* | attested | 4.15 | 1 | 74 | ʃw | *shwossip* | marginal | 3 | 0 |
| 27 | gw | *gweebid* | marginal | 2.85 | 0 | 75 | sm | *smeebid* | attested | 4.05 | 1 |
| 28 | gw | *gwottiff* | marginal | 2.5 | 0 | 76 | sm | *smottiff* | attested | 3.8 | 1 |
| 29 | kl | *kleebid* | attested | 3.95 | 1 | 77 | sn | *sneegiff* | attested | 3.5 | 0 |
| 30 | kl | *klossip* | attested | 4.05 | 1 | 78 | sn | *snossip* | attested | 4.45 | 0 |
| 31 | km | *kmeppid* | unattested | 1.3 | 0 | 79 | sw | *sweegiff* | attested | 3 | 0 |
| 32 | km | *kmossip* | unattested | 1.85 | 0 | 80 | sw | *swezzig* | attested | 3.4 | 0 |
| 33 | kɹ | *kreebid* | attested | 4.25 | 1 | 81 | θw | *thweppid* | attested | 2.2 | 0 |
| 34 | kɹ | *krezzig* | attested | 3.3 | 1 | 82 | θw | *thwossip* | attested | 2.65 | 0 |
| 35 | kw | *kweebid* | attested | 3.5 | 0 | 83 | tl | *tleebid* | unattested | 1.75 | 0 |
| 36 | kw | *kwottiff* | attested | 2.5 | 0 | 84 | tl | *tlottiff* | unattested | 1.84 | 0 |
| 37 | lm | *lmeebid* | unattested | 1.5 | 0 | 85 | tɹ | *trezzig* | attested | 4.05 | 1 |
| 38 | lm | *lmottiff* | unattested | 1.3 | 0 | 86 | tɹ | *trossip* | attested | 5 | 1 |
| 39 | ln | *lneegiff* | unattested | 1.45 | 0 | 87 | tw | *tweegiff* | marginal | 3.4 | 0 |
| 40 | ln | *lnezzig* | unattested | 1.45 | 0 | 88 | tw | *twossip* | marginal | 3.5 | 0 |
| 41 | lt | *ltezzig* | unattested | 1.55 | 0 | 89 | vl | *vleppid* | marginal | 2.3 | 0 |
| 42 | lt | *ltottiff* | unattested | 1.4 | 0 | 90 | vl | *vlossip* | marginal | 3.7 | 0 |
| 43 | ml | *mleppid* | unattested | 1.5 | 0 | 91 | vɹ | *vreebid* | marginal | 2.7 | 0 |
| 44 | ml | *mlossip* | unattested | 1.8 | 0 | 92 | vɹ | *vrezzig* | marginal | 2.3 | 0 |
| 45 | mɹ | *mreegiff* | unattested | 2 | 0 | 93 | vw | *vweegiff* | unattested | 1.7 | 0 |
| 46 | mɹ | *mrottiff* | unattested | 1.7 | 0 | 94 | vw | *vwottiff* | unattested | 1.55 | 0 |
| 47 | nl | *nleebid* | unattested | 1.85 | 0 | 95 | zɹ | *zreppid* | unattested | 2.05 | 0 |
| 48 | nl | *nlezzig* | unattested | 1.35 | 0 | 96 | zɹ | *zrossip* | unattested | 2.1 | 0 |

Table 26: The predicted judgment and the corresponding data in English nonce words; learned from the exception-free learning sample; the predictions deviated from Table 20 are highlighted.

In summary, the proposed learner successfully learns a categorical phonotactic grammar from naturalistic learning samples of English onsets. The learned grammar predicts the categorical judgment of onsets and gradient acceptability of nonce words. The correlation between predicted judgment and human judgment is highly significant. The performance of the Neo-Trubetzkoyan learner in predicting gradient acceptability is on par with the MaxEnt learner.

# 6 Discussion

This section discusses topics that fall beyond the scope of this paper, but merit future studies.

## 6.1 Towards learning-theoretic phonology

The current study subsumes approaches in the growing body of research on phonological learning (Tesar and Smolensky, 2000; Hayes and Wilson, 2008; Gorman, 2013; Heinz, 2010; Tesar, 2014; Jardine, 2016; Gouskova and Gallagher, 2020; Dai and Futrell, 2021). Hayes and Wilson (2008) first introduced the term *learning-theoretic phonology* to describe "a theory whose overall architecture recapitulates the incremental process through which phonological knowledge is acquired". The current study defines learning-theoretic phonology as a research paradigm that probes the nature of phonological grammar by modeling the learning procedure and proposes a standard workflow for learning-theoretic phonology as follows:

1. *Analysis*: Define the learning problem; analyze the conditions for the success/failure of learning (Learnability).

2. *Learning algorithm*: Propose and implement a mathematically grounded computational model for the learning procedure of phonological generalizations;

3. *Simulation*: Train the model on a learning sample consisting only of positive evidence;

4. *Evaluation*: Evaluate the simulation result's prediction on test data consisting of labeled nonce words, preferably verified by behavioral experiments.

This falsifiable research paradigm synthesizes approaches in formal language theory (Heinz, 2007), computational modeling (Hayes and Wilson, 2008; Jarosz, 2019), and the theory of language acquisition (Clark and Lappin, 2010; Lidz and Gagliardi, 2015). Instead of postulating phonological rules/constraints as in standard phonological analysis (Hayes, 2011), learning-theoretic phonology infers a phonological grammar with *minimum* assumptions about hypothesis space, which is usually a specific class of formal languages such as all bigram or trigram constraints (Hayes and Wilson, 2008; Heinz and Rogers, 2010; Wilson and Gallagher, 2018; Gouskova and Gallagher, 2020). In addition, the learning algorithm must state whether the grammar is probabilistic or categorical. The hypothesized learning algorithm is further examined in the simulation of learning from realistic corpus data and evaluated against test data from behavioral experiments.

## 6.2 Comparison with MaxEnt

Hayes and Wilson (2008)'s MaxEnt learner is different from Goldwater and Johnson (2003)'s proposal that learns the constraint weights for a *prespecified* constraint set. Neo-Trubetzkoyan and Hayes and Wilson (2008)'s MaxEnt learner both take up a more challenging task of *directly* inducing phonotactic constraints from corpus data.

Hayes and Wilson (2008)'s MaxEnt learner utilizes the $O/E$ criterion as a heuristic, in which $E$ is obtained with respect to cumulative constraint violations. In contrast, constraint violations in the Neo-Trubetzkoyan learner are noncumulative, and the hypothesis grammar only assigns categorical values such as 0 and 1.

Consider the toy example with the target grammar $\mathcal{T} = \{$*CC$\}$, Table 27 shows the tableaux of a Neo-Trubetzkoyan hypothesis grammar {*CC} and a MaxEnt grammar (Hayes and Wilson, 2008) that penalizes *CC by assigning a higher weight 10 than other constraints. In the Neo-Trubetzkoyan hypothesis grammar, *VC, *CV, *VV are absent and inactive. Any string that violates *CC is ungrammatical; this helps to distinguish ungrammatical exceptions from grammatical words.

| | *CC | $g$ |
|---|---|---|
| CCC | 1 | 0 |
| CCV | 1 | 0 |
| VCC | 1 | 0 |
| CVC | 0 | 1 |

| | 10 | 0 | 0 | 0 | | |
|---|---|---|---|---|---|---|
| | *CC | *CV | *VV | *VC | $h$ | $P^*$ |
| CCC | 2 | 0 | 0 | 0 | 20 | $2.06 \times 10^{-9}$ |
| CCV | 1 | 1 | 0 | 0 | 10 | $4.54 \times 10^{-5}$ |
| VCC | 1 | 1 | 0 | 0 | 10 | $4.54 \times 10^{-5}$ |
| CVC | 0 | 1 | 0 | 1 | 0 | 1 |

Table 27: Neo-Trubetzkoyan hypothesis grammar (left) vs. MaxEnt Grammar (right)

Consider the toy example in §3, the input of the MaxEnt learner is a learning sample $S =$ {CVC, CVV, VVC, VVV, VCV, CCV} where the exception *CCV violates *CC (or *[+cons][+cons]) in the target grammar. The MaxEnt learner initializes an empty hypothesis grammar $G$, then selects natural class-based constraints from the hypothesis space, including *[+cons][+cons], *[+cons]⋉, and *⋊[+cons]. The learner learns the parameter weights of the constraints through the maximization of the likelihood/entropy of the observed learning sample $S$. The output is a MaxEnt grammar consisting of weighted constraints, as shown in Table 28.

| | 1.026 | 0.41 | 0.32 | | |
|---|---|---|---|---|---|
| | *[+cons][+cons] | *[+cons]⋉ | *⋊[-cons] | $h$ | $P^*$ |
| C C | 1 | 1 | 0 | 1.436 | 0.238 |
| C C V | 1 | 0 | 0 | 1.026 | 0.358 |
| V C C | 1 | 1 | 1 | 1.756 | 0.173 |
| V V C | 0 | 1 | 1 | 0.73 | 0.482 |
| V V V | 0 | 0 | 1 | 0.32 | 0.726 |
| C V C V | 0 | 0 | 0 | 0 | 1 |
| V C V C | 0 | 1 | 1 | 0.73 | 0.482 |

Table 28: Learned MaxEnt grammar

Although the learned MaxEnt grammar successfully induced *[+cons][+cons] in the target grammar, it under-penalizes exceptions by assigning nonzero MaxEnt values to strings that violate *CC (CC: 0.238; CCV: 0.358; VCC: 0.173), all of which would receive zero scores in the Neo-Trubetzkoyan learner.

Instead of excluding them, the MaxEnt learner introduces the exceptions to the hypothesis grammar: while maximizing the likelihood of observed learning sample, the MaxEnt learner also maximizes the likelihood of exceptions ("exotic" words per Hayes and Wilson, 2008) such as *CC. The fact that the *CC sequence is observed in the learning sample prevents the MaxEnt learner from assigning a very high penalty to *CC, which could result in a lower score for strings with the *CC sequence. Hayes and Wilson (2008) also reported that their learner obtained "slightly less accurate results from the exceptionful corpora" (P.395). In contrast, the Neo-Trubetzkoyan is capable of removing exceptions from the learning sample following every update of the hypothesis grammar. The behaviors and learning results of both learning algorithms are further compared through the case studies below.

## 6.3 Experimental data and gradience

The current paper evaluated the learning result of English phonotactics in experimental data (Daland et al., 2011). The test data of Turkish nonce words are based on previous documentation (Lees, 1966; Gorman, 2013). Although this is standard practice when experimental data is unavailable (Hayes and Wilson, 2008; Gouskova and Stanton, 2021), future experimental work is required to verify the categorical judgments of nonce words in the test data used in the Turkish case study.

Gradient acceptability judgments in behavioral data challenge phonological learning. An argument for a probabilistic learning model is that probabilistic grammars inherently predict probabilistic distribution, which matches gradient judgments (Clark and Lappin, 2010). However, probabilistic grammar is not a necessary condition for deriving gradient judgments. The categorical grammar in the current proposal can predict a distribution that correlates significantly with gradient acceptability in English. Moreover, although the current proposal utilizes binary categorical grammar, one may also propose a categorical grammar that assigns nonbinary discrete values to words, for example, 1-7 on the Likert scale.

It should be noted that categorical and gradient acceptability judgments must *both* be taken into account when evaluating learning models, because gradient acceptabilities collected in rating tasks could be induced as a task effect (Armstrong et al., 1983; Gorman, 2013). Armstrong et al. (1983) shows that participants rate some odd numbers *odder* than others in rating tasks, challenging the methods used to collect gradient acceptability. A detailed discussion on this matter can be found in Gorman (2013, Chapter 2).

Moreover, rating tasks introduce inconsistent variations. For example, Table 29 shows part of the behavioral data and the predicted judgments of the Neo-Trubetzkoyan learner in Table 23. Sequences with *eppid* receive a higher score than those with *ossip* when paired with /rn/ ($1.65 > 1.45$) but a lower score when paired with /vl/ ($2.3 < 3.7$). One may argue that the /no/ transition in *rnossip* is less well-formed than /ne/ in *rneppid*. However, #77 and #78 exclude this possibility. This inconsistent variation is eliminated in the predicted score $g$ of the Neo-Trubetzkoyan learner.

| # | onset | orthography | attestedness | Likert | $g$ |
|---|---|---|---|---|---|
| 63 | rn | *rneppid* | unattested | 1.65 | 0 |
| 64 | rn | *rnossip* | unattested | 1.45 | 0 |
| 89 | vl | *vleppid* | marginal | 2.3 | 0 |
| 90 | vl | *vlossip* | marginal | 3.7 | 0 |
| 77 | sn | *sneegiff* | attested | 3.5 | 1 |
| 78 | sn | *snossip* | attested | 4.45 | 1 |

Table 29: Variations in Likert scale

In summary, it is possible to incorporate gradience into categorical approaches. While accounting for gradient judgments is an important task, categorical judgments should also be accounted for when evaluating a learning model because gradient judgments from rating tasks might result from task effects and result in inconsistent variations.

## 6.4 Accidental gaps

Accidental gaps are unattested but grammatical sequences that "fall beyond the training space" (Berent et al., 2012). This contradicts the assumption that all unseen data are ungrammatical. For example, in a toy learning sample where an accidental gap ʃʃʃ that is removed {sss, ssʃ, sʃʃ, sʃs, ʃʃʃ, ʃʃs, ʃss, ʃsʃ}, accidental gaps can mislead the learner to conclude that *ʃʃ should be added to the hypothesis grammar. This is because the gap is expected but not observed, which makes $O < E$.

Rawski (2021, Chapter 4) proposed a Bottom-Up Factor Inference Algorithm (BUFIA) for this problem in categorical learning models. BUFIA is based on the insight that the hypothesis space of constraints ($k$-factors) is partially ordered. This allows the search algorithm to go from general to specific and from a smaller $k$-factor to a larger $k$-facotr. BUFIA can avoid reifying accidental gaps because it starts from smaller $k$-factors that are surface true. For example, BUFIA will learn that ʃʃ is acceptable, although a larger $k$-factor ʃʃʃ is missing. Rawski (2021) shows that

BUFIA successfully induced phonotactic constraints in English onsets and Quechua data (Wilson and Gallagher, 2018) that contain accidental gaps. A future direction for the current paper is to integrate Rawski (2021)'s solution to the Neo-Trubetzkoyan learner.

## 6.5 Alternative constraint selection criteria

The current paper proposes a "categorical grammar + constraint selection criterion" approach that learns categorical phonotactic grammar in the presence of exceptions. There are several alternative constraint selection criteria, including Gain (Della Pietra et al., 1997; Berent et al., 2012), Tolerance Principle (Yang, 2016), and the traditional $O/E$ equation (Pierrehumbert, 1993, 2001; Frisch et al., 2004). Gain criterion requires a probability distribution predicted by probabilistic grammars and is incompatible with categorical grammars. The Tolerance Principle requires a well-defined scope to which a rule/constraint can be applied, which is unsuitable for phonotactic models. Before discussing the details, it is worth emphasizing that the computation of $E$ in the Neo-Trubetzkoyan learner is completely different from the traditional $O/E$ equation.[5]

Pierrehumbert (2001)'s $O/E$ has been used to measure the degree of attestedness/representation and predicts gradient acceptability of given 2-factors (Pierrehumbert, 1993; Frisch et al., 2004; Coetzee and Pater, 2008; Gouskova and Stanton, 2021). However, Pierrehumbert (2001)'s equation has been criticized for its mathematical flaws and its inability to model the constraint *strength* in gradient phonotactic grammar (Wilson and Obdeyn, 2009; Stanton and Stanton, 2022; Wilson, 2022)[6].

Moreover, Pierrehumbert (1993)'s equation has been extended to discover phonotactic consonants (Ozburn and Kochetov, 2018; Danis, 2019). Unlike the iterative mechanism in the Neo-Trubetzkoyan learner, all 2-factors with $O/E < 1$ are considered constraints at once (Frisch et al., 2004). However, Pierrehumbert (1993)'s static measure is unsuitable for phonotactic learning because the expected frequency depends on the updated hypothesis grammar. This issue causes the phonotactic grammar discovered by the criterion $O/E < 1$ deviate from the underlying target grammar, even in toy examples. Consider the toy example in §3, where the target grammar $\mathcal{T} = \{$*CC$\}$ underlies the observed sample $S =$ {CVC, CVV, VVC, VVV, VCV, CCV} with an exception *CCV. Table 30 shows the calculated $O/E$ based on Pierrehumbert (1993)'s equation.

[5]The debate over Pierrehumbert (1993)'s $O/E$ equation (Wilson and Obdeyn, 2009; Stanton and Stanton, 2022; Wilson, 2022) is irrelevant to the underlying idea of OE comparison rooted in Trubetzkoy (1939) that has inspired the Neo-Trubetzkoyan learner.

[6]Pierrehumbert (2001) assumes that given a sequence $xy$, the probability that the first symbol appears at the first position $\Pr(x+)$ is independent of the second symbol appearing at the second position $\Pr(+y)$. This assumption is true if the hypothesis grammar remains empty, which is counterfactual if any constraint is added to the grammar (Wilson and Obdeyn, 2009).

| | $O$ | $E$ | $O/E$ |
|---|---|---|---|
| *VV | 3 | $42/11 \approx 3.818$ | 0.79 |
| *VC | 3 | $24/11 \approx 2.181$ | 1.375 |
| *CV | 4 | $35/11 \approx 3.181$ | 1.257 |
| *CC | 1 | $20/11 \approx 1.818$ | 0.55 |

Table 30: The calculated $O/E$ based on Pierrehumbert (1993)'s equation

If $O/E < 1$ indicates the cooccurrence restriction, the learned target grammar should be {*VV, *CC} as both constraints receive $O/E < 1$. However, the learned grammar {*VV, *CC} contradicts the actual target grammar $\mathcal{T} = \{$*CC$\}$.

To summarize, the alternative criteria are not suitable for the current proposal of the phonotactic learner. However, discovering new criteria is a future direction to improve the current proposal.

### 6.6 Other phonological patterns

Although the current paper focuses on nonlocal phonotactics of Turkish vowels and local phonotactics of English onsets, the Neo-Trubetzkoyan learner can also be applied to other phenomena as long as the learner has the correct hypothesis space. A future direction would be to apply the Neo-Trubetzkoyan learner to nonlocal phonological patterns, which can also be characterized in (Tier-based) Strictly 2-Local languages, in other languages such as Quechua (Gouskova and Stanton, 2021, nonlocal laryngeal phonotactics), Hungarian (Hayes and Londe, 2006, vowel harmony), and Arabic (Frisch et al., 2004, nonlocal OCP). It is also possible to extend the learning model to other local phonotactics in Polish word-initial onsets (Kostyszyn and Heinz, 2022; Jarosz and Rysling, 2017), which is similar to the English phonotactics studied in the current paper.

## 7 Conclusion

Lexicalized exceptions constitute a significant source of noise in phonological acquisition. In a positive evidence-only setting, it is common to treat exceptions using indirect negative evidence from distributional information (Clark and Lappin, 2010). Most distribution-sensitive models assume a probabilistic grammar that evaluates the grammaticality of words by their predicted likelihood (Hayes and Wilson, 2008). However, a probabilistic grammar conflates all words into the same spectrum of probability and grammaticality. As a result, short ungrammatical exceptions become more 'grammatical' than longer grammatical words with lower probabilities. This is problematic because it blurs the boundary between exceptions and grammatical words.

The current study proposes the Neo-Trubetzkoyan learner that infers a categorical grammar of phonotactics in the presence of exceptions. This learning model successfully learns from naturalistic corpus data from English and Turkish and predicts categorical and gradient acceptabilities in nonce words. This is possible because the Neo-Trubetzkoyan learner filters out exceptions with

respect to type frequency. The "categorical grammar + statistical criterion" approach provides an explicit demarcation of exceptions and grammatical words, eliminating the need for a special status of exceptions in a probabilistic grammar. This proposal not only provides a compelling alternative to probabilistic approaches but also sheds light on the essence of the long-standing problem of phonotactic learning in the presence of exceptions.

# 8 Appendix

## 8.1 Neo-Trubetzkoyan learner: formal algorithm

The overall learning algorithm is formalized as follows:

**Algorithm 1**: Neo-Trubetzkoyan algorithm

**input** : Training sample $S$
**output**: hypothesis grammar $G$
initialization: $G \leftarrow \{\}, Q \leftarrow \{\}, \textsc{Con} \leftarrow \{C_1, \ldots, C_n\}$, min_difference $\leftarrow 0$;
**do**
  $O \leftarrow$compute_O(Con, $S$);
  $G \leftarrow$ remove_gap($G$);
  $E \leftarrow$approximate_E(Con, $G$, $S$);
  **for** $C \in$ Con **do**
    difference $\leftarrow O[C] - E[C]$;
    $Q \leftarrow Q \cup \{(C, \text{difference})\}$;
    **if** difference $<$ min_difference **then**
      min_difference $\leftarrow$ difference
  **for** $(C, \text{difference}) \in Q$ **do**
    **if** min_difference = difference **then**
      $G \leftarrow G \cup \{C\}$;
      Con $\leftarrow$ Con $- \{C\}$;
      **for** $s \in S$ **do**
        **if** $g(s, G) = 0$ **then**
          $S \leftarrow S - \{s\}$;
**while** $Q \neq \{\}$;

The algorithm for computing $O$ is formalized as follows:

**Algorithm 2:** COMPUTE_O

**input** : The set of constraints Con $= \{C_1, \ldots, C_n\}$ and the learning sample $S$
**output**: A dictionary of constraint-value mappings
initialization: $O \leftarrow \{C_1 : 0, \ldots, C_n : 0\}$ ;
**for** $C \in$ Con **do**
  **for** $s \in S$ **do**
    **if** $C \in \texttt{factor}(s, 2)$ **then**
      $O[C] \leftarrow O[C] + 1$;

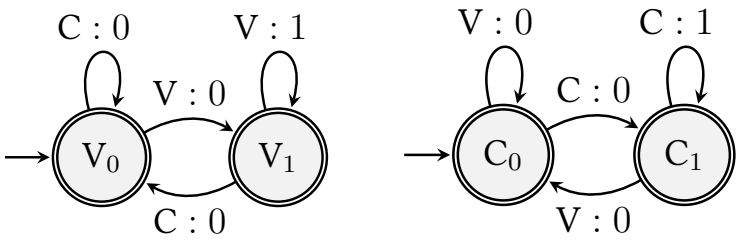
The algorithm for approximating $E$ is formalized as follows:

**Algorithm 3:** APPROXIMATE_E

**input** : The set of constraints Con $= \{C_1, \ldots, C_n\}$, the hypothesis grammar $G$, and the learning sample $S$
**output**: A dictionary of constraint-value mappings
initialization: $E \leftarrow \{C_1 : 0, \ldots, C_n : 0\}$, WFA $\mathcal{M}$ for $G$;
**for** $\ell \in \{1, \ldots, \ell_{\max}\}$ **do**
  initialize $\mathcal{B}_\ell$;
  $\mathcal{N}_\ell \leftarrow \text{COMPOSE}(\mathcal{B}_\ell, \mathcal{M})$;
  **for** $C \in$ Con **do**
    **if** $C \notin G'$ **then**
      $\mathcal{M}' \leftarrow \text{COMPOSE}(\mathcal{M}, \mathcal{M}_C)$;
      $\mathcal{N}'_\ell \leftarrow \text{COMPOSE}(\mathcal{B}_\ell, \mathcal{M}')$;
      $E[C] \leftarrow |S_\ell| \times (1 - (Z(\mathcal{N}'_\ell)/Z(\mathcal{N}_\ell)))$
    **else**
      $E[C] \leftarrow 0$

## 8.2 Weighted Finite-state Automata

$k$-factors, as a subregular language (Rogers and Pullum, 2011), can be characterized by deterministic Weighted Finite-state Acceptors (WFAs), which is essential for the computation of expected frequency in the current proposal. Weighted finite-state acceptors can represent possible strings in a formal language and encode phonotactic grammars. For the convenience of discussion, I assign $w = 0$ for *allowed* transitions and $w = 1$ for *penalized* transitions. WFAs of $\{C_1 : {}^*\text{VV}\}$ and $\{C_2 : {}^*\text{CC}\}$ in Figure 8 For example, in $\mathcal{M}_1$, the transition from state $\text{V}_1$ after accepting V is penalized ($w = 1$), which encodes the constraint *VV. In the intersected WFA $\mathcal{M}$, after accepting V, the transition of V from state V also receives penalty $w = 1$.




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

Figure 8: $\mathcal{M}_1$ for $\{C_1 : {}^*\text{VV}\}$ and $\mathcal{M}_2$ for $\{C_2 : {}^*\text{CC}\}$

The *composition* of WFAs (for a detailed definition, see Mohri et al., 2002, P.6) can be used to update hypothesis grammar. $\mathcal{M} = \mathcal{M}_1 \circ \mathcal{M}_2 \circ \ldots \circ \mathcal{M}_n$ is the composition of WFAs that embodies the combination of individual constraints in the hypothesis grammar $G = \{C_1, \ldots, C_n\}$. In the learning procedure, every new constraint $C$ corresponds to a WFA $\mathcal{M}_C$, and $\mathcal{M}' = \mathcal{M} \circ \mathcal{M}_C$ is the new WFA for the updated hypothesis grammar $G' = G \cup \{C\}$. Consider $\mathcal{M}_1$ the WFA of the original grammar {*VV}, and $\mathcal{M}_2$ encodes a newly added constraint {*CC}. Figure 9 shows the composition of $\mathcal{M}_1$ and $\mathcal{M}_2$ that correspond to the grammar {*CC, *VV}. In this paper, the transition weights on the WFAs represent negative log probabilities.

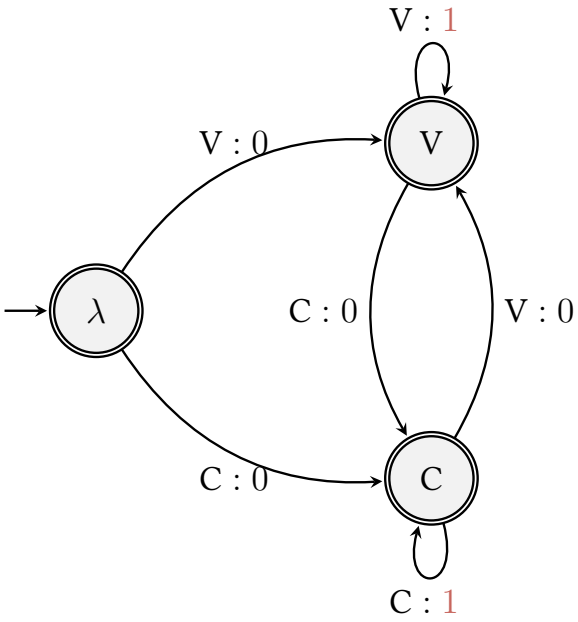



Figure 9: The minimized intersected $\mathcal{M}_1$ and $\mathcal{M}_2$

A braid $\mathcal{B}_\ell$ is a WFA that accepts $\Sigma^\ell$. As shown in Figure 10, when $\ell = 3$, only the states following the paths of exactly three transitions $\text{C}_3$ and $\text{V}_3$ are accepting states in $\mathcal{B}_3$. the state indices indicate the symbol from the last transition and the number of previous transitions. For

example, $V_3$ can only be reached after three transitions from the starting state $\lambda$, and immediately after the machine reads the symbol V.

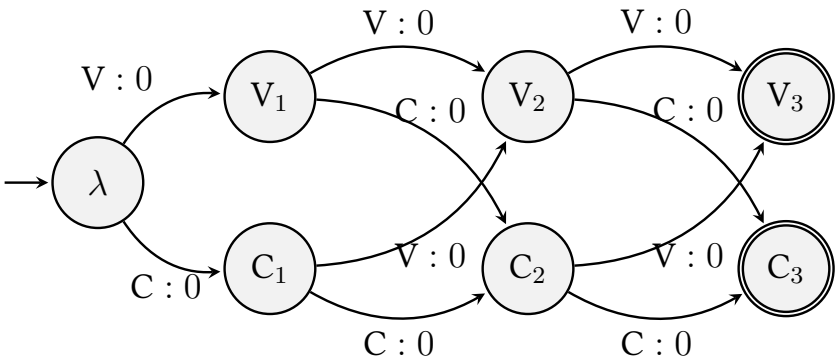



Figure 10: $\mathcal{B}_3$ that accepts all possible 3 length strings

The composition of WFAs can also be used to represent possible strings of certain lengths for a given grammar. As illustrated in Figure 11, $\mathcal{N}_\ell = \mathcal{B}_\ell \circ \mathcal{M}$ is the composition of $\mathcal{B}_\ell$ and the WFA $\mathcal{M}$, which accepts all possible strings $s$ of length $\ell$. In $\mathcal{N}_3$, the paths of allowed strings receive $w = 0$ on every transition, such as $\lambda \xrightarrow{\text{V:0}} \text{V}_1 \xrightarrow{\text{C:0}} \text{C}_2 \xrightarrow{\text{V:0}} \text{V}_3$ (VCV), while the path of a penalized string receives at least one $w = 1$, such as $\lambda \xrightarrow{\text{V:0}} \text{V}_1 \xrightarrow{\text{C:0}} \text{C}_2 \xrightarrow{\text{C:1}} \text{C}_3$. This means the path of VCC is penalized by the current hypothesis grammar encoded in the WFA $\mathcal{N}_3$.

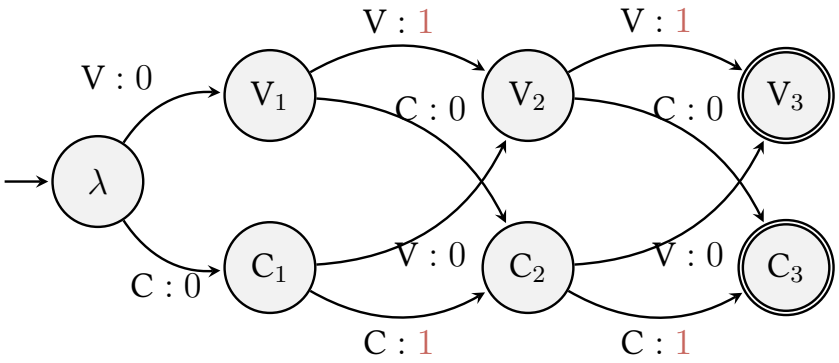



Figure 11: $\mathcal{N}_3 = \mathcal{B}_3 \circ \mathcal{M}$

# References


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

Albright, A. and Hayes, B. (2003). Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, 2(4):343–370.

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13(3):263–308.

Berent, I., Wilson, C., Marcus, G. F., and Bemis, D. K. (2012). On the role of variables in phonology: Remarks on hayes and wilson 2008. *Linguistic inquiry*, 43(1):97–119.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Chomsky, N. and Halle, M. (1965). Some controversial questions in phonological theory. *Journal of linguistics*, 1(2):97–138.

Clark, A. and Lappin, S. (2009). Another look at indirect negative evidence. In *Proceedings of the EACL 2009 workshop on cognitive aspects of computational language acquisition*, pages 26–33.

Clark, A. and Lappin, S. (2010). *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Clements, G. N., Sezer, E., et al. (1982). Vowel and consonant disharmony in turkish. *The structure of phonological representations*, 2:213–255.

Coetzee, A. W. and Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in muna and arabic. *Natural Language & Linguistic Theory*, 26(2):289–337.

Dai, H. and Futrell, R. (2021). Simple induction of (deterministic) probabilistic finite-state automata for phonotactics by stochastic gradient descent. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 167–176. Association for Computational Linguistics.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., and Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, 28(2):197–234.

Danis, N. (2019). Long-distance major place harmony. *Phonology*, 36(4):573–604.

Della Pietra, S., Della Pietra, V., and Lafferty, J. (1997). Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence*, 19(4):380–393.

Frisch, S. A., Pierrehumbert, J. B., and Broe, M. B. (2004). Similarity avoidance and the ocp. *Natural language & linguistic theory*, 22(1):179–228.

Gallagher, G., Gouskova, M., and Rios, G. C. (2019). Phonotactic restrictions and morphology in aymara. *Glossa: a journal of general linguistics*, 4(1).

Göksel, A. and Kerslake, C. (2004). *Turkish: A comprehensive grammar*. Routledge.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Goldwater, S. and Johnson, M. (2003). Learning ot constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, volume 111120.

Gorman, K. (2013). *Generative Phonotactics*. PhD thesis, University of Pennsylvania.

Gorman, K. (2016). Pynini: A python library for weighted finite-state grammar compilation. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 75–80.

Gouskova, M. and Gallagher, G. (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory*, pages 1–40.

Gouskova, M. and Stanton, J. (2021). Learning complex segments. *Language*, 97(1):151–193.

Hayes, B. (2011). *Introductory phonology*, volume 32. John Wiley & Sons.

Hayes, B. and Londe, Z. C. (2006). Stochastic phonological knowledge: The case of hungarian vowel harmony. *Phonology*, 23(1):59–104.

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

Heinz, J. (2007). *The inductive learning of phonotactic patterns*. PhD thesis, PhD dissertation, University of California, Los Angeles.

Heinz, J. (2010). Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4):623–661.

Heinz, J. (2011a). Computational phonology–part i: Foundations. *Language and Linguistics Compass*, 5(4):140–152.

Heinz, J. (2011b). Computational phonology–part ii: Grammars, learning, and the future. *Language and Linguistics Compass*, 5(4):153–168.

Heinz, J. (2018). The computational nature of phonological generalizations. *Phonological Typology, Phonetics and Phonology*, pages 126–195.

Heinz, J. and Idsardi, W. J. (2017). Computational phonology today. *Phonology*, 34(2):211–219.

Heinz, J., Rawal, C., and Tanner, H. G. (2011). Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 58–64. Association for Computational Linguistics.

Heinz, J. and Rogers, J. (2010). Estimating strictly piecewise distributions. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 886–896.

Hughto, C., Lamont, A., Prickett, B., and Jarosz, G. (2019). Learning exceptionality and variation with lexically scaled maxent. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 91–101.

Hyman, L. M. (1975). *Phonology: Theory and Analysis*. Holt, Rinehart & Winston.

Inkelas, S., Küntay, A., Orgun, C. O., and Sprouse, R. (2000). Turkish electronic living lexicon (TELL): A lexical database. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Jardine, A. (2016). Learning tiers for long-distance phonotactics. In *Proceedings of the 6th conference on generative approaches to language acquisition North America (GALANA 2015)*, pages 60–72.

Jardine, A. and Heinz, J. (2016). Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics*, 4:87–98.

Jarosz, G. (2019). Computational modeling of phonological learning. *Annual Review of Linguistics*, 5:67–90.

Jarosz, G. and Rysling, A. (2017). Sonority sequencing in polish: The combined roles of prior bias & experience. In *Proceedings of the annual meetings on phonology*, volume 4.

Jelinek, F. (1998). *Statistical methods for speech recognition*. MIT press.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kostyszyn, K. and Heinz, J. (2022). Categorical account of gradient acceptability of word-initial polish onsets. In *Proceedings of the Annual Meetings on Phonology*, volume 9.

Lau, J. H., Clark, A., and Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Lees, R. B. (1966). On the interpretation of a turkish vowel alternation. *Anthropological Linguistics*, pages 32–39.

Lidz, J. and Gagliardi, A. (2015). How nature meets nurture: Universal grammar and statistical learning. *Annual Review of Linguistics*, 1(1):333–353.

Linzen, T., Kasyanenko, S., and Gouskova, M. (2013). Lexical and phonological variation in russian prepositions. *Phonology*, 30(3):453–515.

Mohri, M. (2002). Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350.

Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Moore-Cantwell, C. and Pater, J. (2016). Gradient exceptionality in maximum entropy grammar with lexically specific constraints. *Catalan Journal of Linguistics*, 15:53–66.

Osherson, D., Stob, M., and Weinstein, S. (1986). *Systems that learn: an introduction to learning theory*. MIT press.

Ozburn, A. and Kochetov, A. (2018). Ejective harmony in lezgian. *Phonology*, 35(3):407–440.

Pater, J. (2000). Non-uniformity in english secondary stress: the role of ranked and lexically specific constraints. *Phonology*, 17(2):237–274.

Pearl, L. and Lidz, J. (2009). When domain-general learning fails and when it succeeds: Identifying the contribution of domain specificity. *Language Learning and Development*, 5(4):235–265.

Pearl, L. S. and Mis, B. (2016). The role of indirect positive evidence in syntactic acquisition: A look at anaphoric" one". *Language*, pages 1–30.

Perkins, L., Feldman, N. H., and Lidz, J. (2022). The power of ignoring: Filtering input for argument structure acquisition. *Cognitive Science*, 46(1):e13080.

Pierrehumbert, J. (1993). Dissimilarity in the arabic verbal roots. In *Proceedings of NELS*, volume 23, pages 367–381. University of Massachusetts Amherst.

Pierrehumbert, J. (2001). Stochastic phonology. *Glot international*, 5(6):195–207.

Prince, A. and Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell.

Rawski, J. (2021). *Structure and Learning in Natural Language*. PhD thesis, State University of New York at Stony Brook.

Regier, T. and Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition*, 93(2):147–155.

Riggle, J. A. (2004). *Generation, recognition, and learning in finite state Optimality Theory*. University of California, Los Angeles.

Rogers, J. and Pullum, G. K. (2011). Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*, 20(3):329–342.

Stanton, J. and Stanton, J. F. (2022). In defense of O/E. *lingbuzz/006391*.

Tesar, B. (2014). *Output-driven phonology: Theory and learning*. Cambridge University Press.

Tesar, B. and Smolensky, P. (2000). *Learnability in optimality theory*. MIT Press.

Trubetzkoy, N. S. (1939). *Grundzüge der phonologie*. Prague: Travaux du cercle linguistique de Prague 7.

Trubetzkoy, N. S. (1969). *Principles of Phonology* (Christiane A. M. Baltaxe, Trans.). University of California Press, Berkeley and Los Angeles.

Wilson, C. (2022). Identifiability, log-linear models, and observed/expected (response to stanton & stanton, 2022). *lingbuzz/006474*.

Wilson, C. and Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: A case study of south bolivian quechua. *Linguistic Inquiry*, 49(3):610–623.

Wilson, C. and Obdeyn, M. (2009). Simplifying subsidiary theory: statistical evidence from arabic, muna, shona, and wargamay. ms. Johns Hopkins University.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

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

Zuraw, K. R. (2000). *Patterned exceptions in phonology*. University of California, Los Angeles.