An Exception-Filtering Approach to Phonotactic Learning

Huteng Dai

Abstract

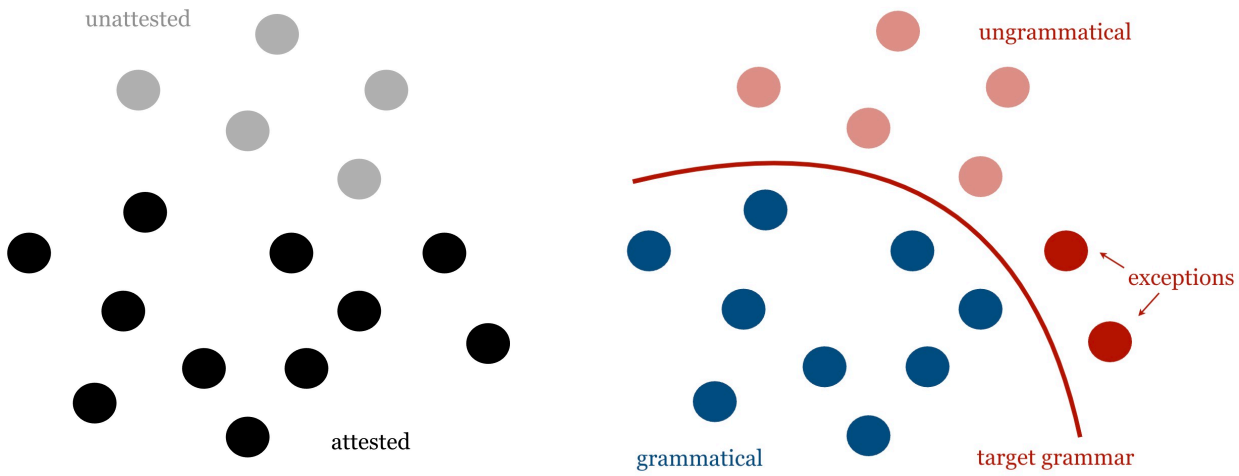
Phonotactic learning has been a fertile ground for research in the field of phonology. However, the challenge of lexical exceptions in phonotactic learning remains largely unexplored. Traditional learning models, which typically assume all observed input data as grammatical, often blur the distinction between lexical exceptions and grammatical words, consequently skewing the learning results. To address this issue, this paper innovates a "categorical grammar + exception-filtering mechanism" approach to iteratively filter out ungrammatical sequences, utilising frequency information from input data. Applied to a variety of naturalistic corpora from English, Polish, and Turkish, this method showed a high correlation with native speakers' acceptability judgments in behavioural experiments, highlighting its capability for handling lexical exceptions.

Keywords: phonotactics, phonological learning, categorical grammar, exceptionality, indirect negative evidence, frequency, onsets, vowel harmony, acceptability, English, Polish, Turkish

1 Introduction

There exist logically infinite potential sound sequences in any given language, yet only some are considered well-formed. This implicit knowledge, which enables native speakers to recognise these well-formed sequences, is known as *phonotactics*. However, phonotactic knowledge does not apply uniformly to the entire lexicon. On the contrary, certain lexical exceptions violate otherwise universally applicable patterns (Guy, 2007). For example, despite the existence of productive vowel harmony patterns (Zimmer, 1969; Kabak, 2011), all disharmonic sequences of Turkish vowels of length two are attested in the Turkish Electronic Living Lexicon (Inkelas et al., 2000), which includes sound sequences articulated by native adult speakers (see §7 for details).

The presence of lexical exceptions poses a challenge to learning models that assumed exception-free input data. This challenge is illustrated in Figures 1a and 1b. Under the positive evidence-only assumption, the learner relies exclusively on unlabelled input data (Marcus, 1993; Clark and Lappin, 2010), denoted by the darker dots in the figures; conversely, lighter dots represent unattested data that are absent from the input. The learning problem is to arrive at a target grammar that can differentiate between grammatical and ungrammatical data, a distinction represented by the curve in Figure 1b.



(a) Potential sound sequences

(b) Target grammar

Figure 1: The learning problem in the presence of exceptions. Darker dots represent attested data, while light dots indicate unattested onsets; inspired by Mohri et al. (2018: 8).

Learning models that assume all attested sound sequences, including lexical exceptions, as part of the grammatical patterns run the risk of building noise into the model. This is a case of "overfitting" in machine learning, in which a model is trained too well on the input data, to the extent that it starts to fit noise, consequently reducing its ability to generalise to unseen data (Mohri et al., 2018). Therefore, the most optimal model does not necessarily fit the input data perfectly; instead, it should filter out or heavily penalise lexical exceptions as perceived noise.

Although exceptionality has been a perennial interest in phonology (Wolf, 2011; Moore-Cantwell and Pater, 2016; Mayer et al., 2022)[1], learning models capable of handling exceptions and based on categorical grammars remain to be developed. Categorical grammars provide clear-cut demarcation between grammatical sequences and lexical exceptions. However, learning models based on categorical grammars are generally considered vulnerable to exceptions in naturalistic corpora, as discussed in Gouskova and Gallagher (2020) (emphasis added):

> "In contrast to our approach, Heinz (2010), Jardine (2016), and Jardine and Heinz (2016) characterise non-local phonology as an idealised problem of searching for unattested substrings. Their learners memorise 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)."

[1]The challenge of exception in phonotactic learning is analogous to that of "Type IV" patterns in Moreton et al. (2017), which can be conceptualised as general patterns that have a single exception. Their learning model took longer to learn Type IV patterns compared to exceptionless patterns, but eventually reached convergence. This difficulty was mirrored in their learning experiment. The author thanks a reviewer for showing this connection.

However, it would be uninsightful to dismiss categorical grammars altogether based on the modest performance of several idealised models, which were designed to explore the mathematical underpinnings of phonological learning, not to handle real-world corpora. Recent developments have both demonstrated promising results using simple categorical phonotactic learning models in naturalistic corpora (Gorman, 2013; Durvasula, 2020; Kostyszyn and Heinz, 2022) and begun to address complex challenges such as accidental gaps (Rawski, 2021).

The current study undertakes a similar endeavour: rooted in Formal Language Theory, it proposes an novel approach to address the problem of exceptions by integrating frequency information from the input data. This proposal draws inspiration from probabilistic approaches, especially the learner in Hayes and Wilson (2008) and traditional O/E criterion (Trubetzkoy, 1939; Pierrehumbert, 1993), and takes the initiative to bridge the gap between the mathematical underpinnings of phonological learning and realistic data, harnessing the potential that categorical grammars can offer. The learning model proposed in the current study demonstrates robust performance in realistic corpora in English, Polish, and Turkish, by learning categorical grammars that closely align with acceptability judgments in behavioural experiments. In particular, the learner successfully acquires nonlocal vowel phonotactics in Turkish, despite the complexity introduced by disharmonic roots and derived forms in highly exceptionful input data.

This paper is structured as follows: §2 outlines the theoretical background and related assumptions; §3 introduces the current proposal—the Exception-Filtering learning algorithm; §4 illustrates the evaluation methods and provides an overview of the three subsequent case studies in English (§5), Polish (§6), and Turkish (§7). §8 discusses topics emerged from the current study and outlines directions for future work.

2 Background

This section outlines the essential concepts, underlying assumptions, and relevant evidence involved in the current proposal.

2.1 Competence-performance Dichotomy

This study assumes three interconnected components involved in phonotactic learning: grammar, lexicon, and performance.[2] Inspired by the *dual-route* model (Pinker and Prince, 1988; Zuraw, 2000; Zuraw et al., 2021)[3], the relationship among these components is visualised in Figure 2.

[2]Hale and Reiss (2008: 18) adopted a nihilistic view, arguing that phonotactics is computationally inert in morphophonological alternations and not a part of grammar (Reiss 2017: §6). However, ample evidence indicates that infants acquire and utilise phonotactics (Jusczyk et al., 1993, 1994; Jusczyk and Aslin, 1995; Archer and Curtin, 2016). Gorman (2013: §1) also demonstrated the internalization of phonotactic constraints in various domains, such as wordlikeness judgements and loanword adaptation. Moreover, the current study upholds the concept of "categorical grammars", which essentially motivated the adoption of the nihilistic view (Reiss 2017: 14; " categorical baby"). Therefore, the question of how categorical phonotactic grammars are learnt is still relevant.

[3]This paper primarily borrows the dual-route model's core concepts of the lexical/non-lexical routes, while ignoring certain controversial topics of this model, such as the analogy mechanism (Pinker and Prince, 1988) and whole-word storage in morphological processing (Lignos and Gorman, 2012; Yang, 2016).

Together, lexicon and grammar form the *competence*, representing the internalised knowledge. The potential interaction between the lexicon and grammar (Ernestus and Baayen, 2003; Martin, 2011) is indicated by the dashed line in Figure 2. Phonotactic learning involves the acquisition of a grammar from primary linguistic data, which is gleaned from the performance of other speakers and is influenced by both grammar and extragrammatical factors.

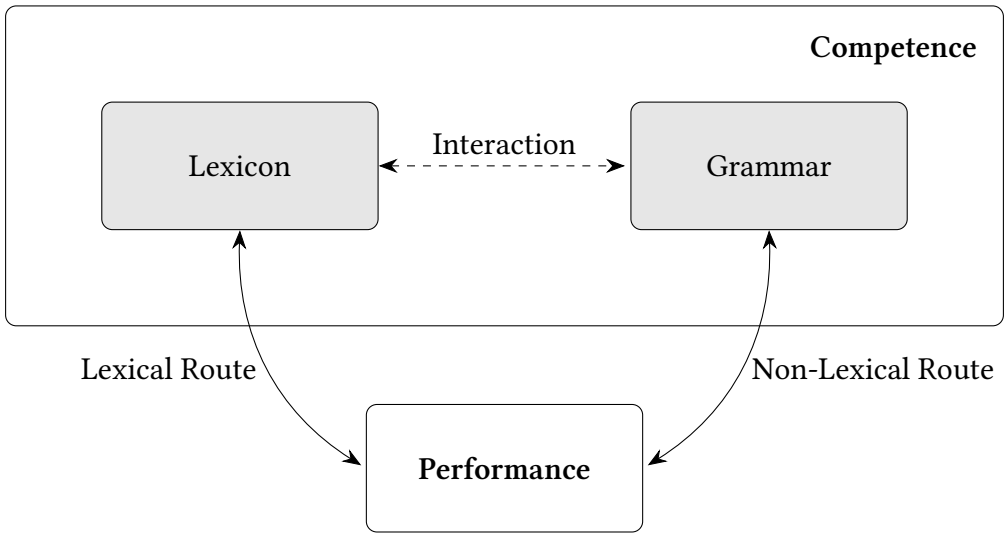


Figure 2: The relationship between lexicon, grammar, and performance in a dual-route model

The lexical route allows the speaker to access the lexicon and evaluate the acceptability of sound sequences, regardless of potential grammar violations. If the lexicon does not contain certain sound sequences, as is the case with nonce words, the speaker instead evaluates their acceptability in the grammar via the non-lexical route, in which grammaticality is predicted based on the grammar. The current study distinguishes between the terms *grammaticality* (or well-formedness) and *acceptability*, which have frequently been conflated in previous research (Hayes and Wilson, 2008; Albright, 2009). In this context, acceptability refers to the judgments made by native speakers on real-world performance, which can be influenced by both grammar and extragrammatical factors, such as processing difficulty, lexical frequency, and similarity (Schütze, 1996, see detailed discussion §8). In contrast, grammaticality refers to the abstract, internalised knowledge represented by the grammar, such as phonotactic constraints in the current paper, independent of any extragrammatical factors, such as frequency information. A sound sequence is deemed grammatical *only if* it adheres strictly to the hypothesis grammar.

Therefore, the relationship between grammaticality and acceptability is not one-to-one: certain ungrammatical forms in the lexicon can be deemed more acceptable than some grammatical forms. A model that perfectly aligns with acceptability, paradoxically, deviates from the grammar. This is not due to its inability to explain acceptability, but rather to its overreach in explanatory power, which is achieved by representing extragrammatical factors in grammar (Kahng and Durvasula 2023: 3).

Acceptability judgments are commonly collected via rating tasks employing a numeric Lik-

ert scale and characterised as "gradient" (non-categorical) in nature (Albright, 2009). Mathematically, individual Likert ratings correspond to multi-level categorical, rather than continuous, values, e.g., 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree, exhibiting considerable individual variability, which are not incompatible with categorical grammars.[4] When averaged over multiple participants, these results can present as gradient values, hinting at the need to incorporate individual variability within a categorical framework (see §8 for a discussion). Furthermore, influenced by task effects, rating tasks can elicit gradient responses even for inherently discrete phenomena, such as the concept of odd and even numbers (Armstrong et al., 1983; Gorman, 2013). Another extragrammatical factor at play in the acceptability judgment is traced back to *auditory illusions*, as shown in Kahng and Durvasula (2023). In light of these considerations, the acceptability judgments reported in previous studies are not incompatible with categorical grammar. Therefore, the current study argues that gradient acceptability judgments collected via numeric rating tasks do not necessitate gradient / probabilistic grammars or negate the possibility of categorical grammars (cf. Coleman and Pierrehumbert, 1997; Hayes and Wilson, 2008).

Although grammaticality should not be conflated with acceptability in behavioural data, the current study assumes that the grammaticality of sound sequences, categorical or probabilistic, is *reflected* in acceptability judgments, and a successful grammar should exhibit a robust correlation between predicted grammaticality and acceptability judgments (Lau et al., 2017).

The current study employs categorical grammars using a discrete set of constraints that simply accept grammatical sequences and reject ungrammatical ones. In contrast, probabilistic grammars, such as Maximum Entropy (MaxEnt) grammars (Hayes and Wilson, 2008), involve constraints along with continuous weights, assigning a probability continuum across all possible sequences. Analogous to probabilistic grammars, grammaticality in categorical grammars is associated with discrete, often binary values, where 0 signifies ungrammatical sequences, and 1 designates grammatical ones. Probabilistic grammars have been noted for their ability to model human sensitivity to frequency information and approximate gradient acceptability judgments (Hayes and Wilson, 2008), while categorical grammars delineate a clear boundary between grammatical and ungrammatical sequences, thus avoiding the potential conflation of grammar with lexicon and other extragrammatical factors.

2.2 Attestedness vs Grammaticality

The term "lexicon" in this paper is strictly employed to refer to the shared lexicon among speakers. The "developing lexicon" of individual learners, while significant, is beyond the scope of this paper. Lexicon was first introduced in Chomsky's *Aspects* as the repository of all known words shared among speakers, including all exceptional, unpredictable features of attested input data (Bloomfield, 1933; Chomsky, 1965; Jackendoff, 2002), while grammar acts as a finite system representing infinite number of grammatical sound sequences (Chomsky and Halle, 1968).

[4]Alternatively, categorical grammar can represent nonbinary discrete contrasts. For example, categorical multilevels, such as 1 (ungrammatical), 3 (marginal), and 5 (grammatical), can be achieved by distributing potential constraints into three distinct subsets of the grammar. Although the current study does not adopt this alternative, such a method could be advantageous for modelling intermediate acceptability judgments.

Consequently, the input data are drawn from the assumed shared lexicon and can include sound sequences that deviate from the grammar.

For convenience in the discussion, consider a hypothesis grammar that consists of constraints {*sf, *bn}. The symbol * is only used to indicate ungrammatical sequences (opposed to unattested). As illustrated in Table 1, attestedness indicates whether a sound sequence occurs in the input data. For example, [brɪk] (as in *brick*) and [*sfiɚ] (**sphere*) are both attested in the English lexicon, while [blɪk] (*blick*) and [*bnɪk] (**bnick*) are not.

	grammatical	**ungrammatical**
attested	[brɪk]	[*sfiɚ]
unattested	[blɪk]	[*bnɪk]

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

In contrast, grammaticality indicates whether phonological representations conform to the hypothesis grammar, which is reflected in acceptability judgments. This discrepancy between attestedness and grammaticality yields accidental gaps (grammatical but unattested) and exceptions (attested but ungrammatical), with this paper particularly emphasising the latter. For example, although both are nonexistent words, *blick* is grammatical while **bnick* is not, as native speakers uniformly reject **bnick* while accepting *blick* (Chomsky and Halle, 1965).

While *blick* is a classic example of accidental gaps, *sphere* is a classic example of lexical exceptions. The onset [sf] rarely occurs in English and is often deemed ungrammatical (Hyman, 1975; Algeo, 1978; Kostyszyn and Heinz, 2022), although the acceptability of specific words such as "sphere" may vary among speakers due to factors such as lexical knowledge (Bailey and Hahn, 2001). Moreover, [sf]-onset nonce words are commonly judged unacceptable, as shown in an experiment conducted by Scholes (1966): 33 English speakers are asked whether a nonce word "is likely to be usable as a word of English." Only 7 participants responded "yes" to the [sf]-onset nonce word [sfid], lower than [blʌŋ] (31 "yes"), and even lower than words with unattested onsets such as [mlʌŋ] (13 "yes"). Other examples of exceptions in English include [sɪksθs] (*sixths*), [-lfθs] (*twelfths*), and [ŋsts] (*angsts*), where attested complex clusters are prone to speech errors, thus indicating the violation of phonotactic constraints within the lexicon (Fromkin, 1973).

Lexical exceptions are commonly observed in loanwords, leading to an evolving lexicon that could incorporate ungrammatical sound sequences from various languages (Kang, 2011). For example, exceptional onsets can be observed in English loanwords, such as [bw] *Bois*, [sr] *sri*, [ʃm] *schmuck*, [ʃl] *schlock*, [ʃt] *shtick*, [zl] *zloty*, and adapted names from different languages, including [vr] *Vradenburg*. An important observation is that all these onsets exhibit low type frequencies in English, according to the CMU Pronouncing Dictionary (Weide et al., 1998, www.speech.cs.cmu.edu/cgi-bin/cmudict). Similar examples have been observed in other languages where putative phonotactic restrictions do not extend to loanwords (Gorman 2013: 6-7). Thus, this paper takes the position that the lexicon is an extensive repository that can contain ungrammatical words according to the hypothesised phonotactic grammar. In turn, the input data drawn from the lexicon can consist of lexical exceptions.

The goal of a phonotactic learner is to select the grammar that distinguishes between grammatical and ungrammatical sequences from unlabelled input data. This problem is challenging in the presence of exceptions because intrusions of ungrammatical sequences can mislead the learner to build exceptional patterns in the hypothesis (Clark and Lappin, 2010). Computationally, a learning model exposed solely to positive evidence struggles to identify the target grammar from the hypothesis space of numerous formal language classes (Gold, 1967; Osherson et al., 1986). This challenge is particularly evident in classes of linguistic interest, such as the (Tier-based) Strictly 2-Local languages utilised in §3. An in-depth review of this issue can be found in Wu and Heinz (2023).

One approach to address the challenge of exceptions utilises an *exception-filtering* mechanism to exclude exceptions while learning categorical grammars. Although such a mechanism was considered challenging to propose (Clark and Lappin 2010: 105), the current study achieves this by leveraging *indirect negative evidence* derived from frequency information (Clark and Lappin, 2009, 2010; Regier and Gahl, 2004; Pearl and Lidz, 2009; Pearl and Mis, 2016; Yang, 2016), specifically from type frequency (Pierrehumbert, 2001a; Albright and Hayes, 2003; Hayes and Londe, 2006; Hayes and Wilson, 2008; Albright, 2009).[5] Indirect negative evidence allows learners to infer grammaticality labels from unseen data, despite the absence of such labels in positive evidence. For example, learners can infer that words with an infrequent onset [sf] are ungrammatical. This is guided by the principle that a sequence that occurs less frequently than expected in the input data is likely ungrammatical. This approach compares observed and expected frequencies, a long-standing method in phonology for identifying illicit patterns (Trubetzkoy, 1939; Pierrehumbert, 1993; Frisch et al., 2004; Hayes and Wilson, 2008).

2.3 Summary

The current study has underscored the tension between competence and performance and clarified the nuanced distinctions between acceptability and grammaticality within a dual-route model. It uses a categorical grammar that distinguishes between grammatical and ungrammatical data. The current study argues that the learning model should correlate the grammaticality scores predicted by the learnt grammar with acceptability judgments and handle lexical exceptions by using an exception-filtering mechanism based on frequency information.

3 Exception-Filtering Phonotactic Learner

This seciton proposes a "categorical grammar + exception-filtering" approach to select a hypothesised categorical grammar (hereafter "hypothesis grammar") from the hypothesis space. This

[5]Lexical exceptions might also exhibit unexpectedly high *token frequencies*. For example, the disharmonic Turkish word [silɑh] "weapon" contradicts the backness harmony pattern, yet has a frequency of 26,658. On the contrary, the grammatical root [sɑpɯk] "pervert" is less common, with only 2,716 occurrences in a Wiki corpus of approximately 100 million words (https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Turkish_WordList_10K). However, previous studies have shown that type frequency yields better results in modelling phonological intuitions (Hayes and Wilson 2008: 395). The current study leaves this alternative strategy for future investigation.

section starts by justifying the concepts and assumptions of the current proposal and then introduces the core learning algorithm in §3.4.

3.1 Segment-based Representation

The current proposal adopts segmental representations derived from the input data, a departure from the prespecified feature representations advocated by previous studies (Hayes and Wilson, 2008; Gouskova and Gallagher, 2020). Segmental representations risk misinterpreting accidental gaps as systematic constraints and may overlook sub-segmental generalisations. As Hayes and Wilson (2008: 401) demonstrated, a feature-based model outperforms a segment-based model in their English case study. The problem of selecting feature-based constraints is beyond the scope of this study, while §8 demonstrates a promising solution.

Moreover, the primary goal of this study is not to build an all-around model of phonotactic learning, but to distill the problem of exceptions to its essence at a computational level (Marr, 1982). In this paper, a segmental approach facilitates the analysis of exceptions tied to segment-based constraints. For example, the presence of [sf] in the word *sphere* explicitly violates a single segmental constraint *sf but could be associated with several feature-based constraints such as *[+sibilant, -voice][+labiodental, -voice] and *[+alveolar][+labiodental]. Moreover, segmental representations can be directly obtained from the input data, independent of any prespecified feature system. Employing segmental representations also significantly narrows down the hypothesis space as discussed below.

3.2 The Structure of Grammars and Hypothesis Space

Phonotactic learning involves selecting a hypothesis grammar (G; a set of constraints) from the hypothesis space (Con; borrowed from OT terminologies). The current study uses a noncumulative, inviolable and unranked categorical grammar, labelling any sequence with nonzero constraint violations as "ungrammatical" and those with zero violations as "grammatical". The current study intentionally departs from the *cumulative effects* suggested in previous experimental work (Coleman and Pierrehumbert, 1997; Breiss, 2020; Kawahara and Breiss, 2021), and primarily investigates whether phonotactic learning of categorical grammars is possible in the presence of exception. One possibility to incorporate cumulativity in the future could involve replacing the grammaticality function with the sum of constraint violations (see also §8).

This structure of grammars, while similar, diverges significantly from the cumulative, violable, and ranked grammar in Optimality Theory (OT; Prince and Smolensky, 1993; Prince and Tesar, 2004). In contrast to OT, the hypothesis grammar in the current proposal is drawn from a highly restrictive hypothesis space.[6] Based on the analytical results of Formal Language Theory (FLT), the current study adopts Tier-based Strictly k-Local (TSL$_k$) languages (Heinz et al., 2011; Jardine and Heinz, 2016; Lambert and Rogers, 2020) as the hypothesis space. In Formal Language Theory, the meanings of "language" deviate from their literal meanings. A language is a set of

[6]For an in-depth discussion on the computational complexity of OT grammars, refer to works such as Ellison (1994); Eisner (1997); Idsardi (2006); Heinz et al. (2009).

strings (e.g., sound sequences) that adhere to its associated grammar, which can be characterised as a set of forbidden structures.

k-factors are substrings of length k. A TSL$_k$ grammar consists of all forbidden k factors on a specific tier, known as TSL$_k$ constraints. The tier, also referred to as a *projection* (Hayes and Wilson, 2008), functions as a targeted subset of the inventory of phonological representations (e.g., segments, consonants, vowels) for constraint evaluation. In the context of local phonotactics, the tier encompasses the full inventory, such as all segments, while in nonlocal phonotactics, it includes only specific segments, such as vowels. For example, as shown in Figure 3, a Turkish word [døviz] "currency" is represented as [øi] on the vowel tier. Nontier segments are ignored during the evaluation of tier-based constraints. Therefore, [døviz] violates a tier-based constraint *øi on the vowel tier. This concept, while similar, is distinct from the traditional feature-based definition in Autosegmental Phonology (Goldsmith, 1976)

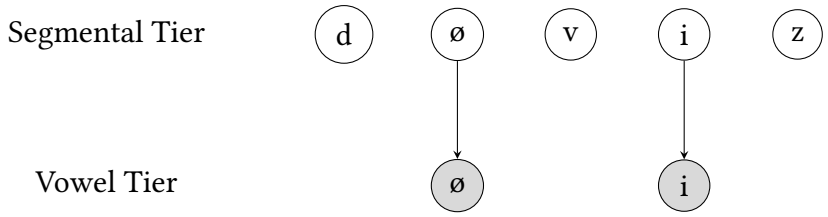


Figure 3: Extraction of vowel tier from the Turkish word [døviz] "currency". The vowel tier contains the vowels in this word, disregarding the non-tier consonants.

A string is labelled as grammatical if it does not contain any forbidden k-factors specified by the grammar; otherwise, the string is considered ungrammatical. This can be formalised by the function $\texttt{factor}(s, k)$, which generates all k-factors of a string s. For example, $\texttt{factor}(\text{CCV}, 2) = \{\text{CC}, \text{CV}\}$, and $\texttt{factor}(\text{CVC}, 2) = \{\text{CV}, \text{VC}\}$. The grammaticality score of a string s under a grammar G, denoted as $g(s, G)$, is defined as follows:

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

For example, consider a grammar $G = \{\text{*CC}\}$, which forbids any strings containing the sequence CC. In this case, the string CCV would be deemed ungrammatical, while the string CVC would be classified as grammatical.

TSL$_k$ languages delineate a formally restrictive yet typologically robust hypothesis space, capturing a range of local and nonlocal phonotactics (Heinz et al., 2011). Specifically, McMullin and Hansson (2019) provides experimental evidence for TSL$_2$ as an appropriate hypothesis space for phonotactic learning, demonstrating that participants in artificial learning experiments were able to learn TSL$_2$ patterns, but struggled with patterns that fall outside the TSL$_2$ class. Formal language-theoretic studies have also demonstrated that this hypothesis space is accompanied by efficient learning properties (Heinz et al., 2011; Jardine and Heinz, 2016; Jardine and McMullin, 2017). This approach has been successfully applied in previous work that span both probabilistic

and categorical approaches (Hayes and Wilson, 2008; Gouskova and Gallagher, 2020; Mayer, 2021; Dai et al., 2023; Heinz, 2007; Jardine and Heinz, 2016).

One of the main challenges of phonotactic learning, as mentioned in Hayes and Wilson (2008: 392), is the rapid growth of the hypothesis space with increasing size of k. In response to this challenge, the current study limits k to two (TSL_2), which is sufficient to capture a large amount of local and nonlocal phonotactic patterns. Although this paper only examines local phonotactics of English and Polish onsets and nonlocal phonotactics of Turkish vowels, the proposed hypothesis space is broadly applicable for suitable domains, extending to phenomena such as nonlocal laryngeal phonotactics in Quechua (Gouskova and Stanton, 2021), Hungarian vowel harmony (Hayes and Londe, 2006), and Arabic OCP-Place patterning (Frisch and Zawaydeh, 2001; Frisch et al., 2004). To summarise, the learner hypothesises a noncumulative, inviolable, and unranked categorical TSL_2 grammar, derived from the hypothesis space of TSL_2 languages.

3.3 Exception-Filtering Mechanism: O/E Criterion

The current study builds on the assumption that that unique ungrammatical sound sequences, including exceptions, are generally observed less frequently than expected, following Hayes and Wilson (2008). This comparison between observed (O) and expected (E) type frequencies embodies the exception-filtering mechanism in the current study (see §8 for alternative criteria) and has been widely applied in identifying phonotactic constraints (Trubetzkoy, 1939; Pierrehumbert, 1993, 2001a; Frisch et al., 2004; Hayes and Wilson, 2008). For instance, the exceptional [sf] sequence would have the same expected type frequency as grammatical sequences like [br] (as in *brick*) if no constraints are present in the current grammar. However, if [sf] only appears in a limited number of words, such as *sphere*, its observed type frequency would be significantly lower than its expected type frequency. This discrepancy allows the learner to infer a *sf constraint and classify the observed *sphere* as a lexical exception. Hayes and Wilson (2008) hypothesised that children possess an innate ability to discern the unique status of exotic items and improved their learning results by excluding these exotic items from input data (Hayes and Wilson, 2008: 427-428). This ability to detect and exclude anomalies aligns closely with the concept of exception-filtering in the current proposal.

The traditional O/E equation proposed by Pierrehumbert (1993) is frequently used for constraint discovery (Pierrehumbert, 2001a; Frisch et al., 2004). However, this equation assumes an empty hypothesis grammar, which becomes inaccurate once any constraint is added, as discussed in Wilson and Obdeyn (2009) and Wilson (2022).

The current criterion O/E draws inspiration from Hayes and Wilson (2008). However, a key divergence lies in the fact that the hypothesis grammar in the current study is noncumulative, leading to distinct calculations O and E. The observed type frequency (O) of a potential constraint C is determined by the count of *unique* strings in the sample that violate C:

$$O[C] = |\{s \in S : C \in \texttt{factor}(s, 2)\}| \tag{2}$$

In a toy sample $S = \{$CVC, CVV, VVC, VVV, VCV, CCV$\}$, $O[$*CC$] = 1$, $O[$*CV$] = 4$, $O[$*VC$] = 3$, $O[$*VV$] = 3$. Here, $O[$*VV$]$ is not 4 because the learner only counts unique strings violated by

the potential constraint, rather than cumulative violations. Moreover, O is updated during the learning process, as the learner filters out lexical exceptions from the input data S every time a new constraint is added to the hypothesis grammar.

The expected type frequency $E[C]$ represents the number of unique strings in the hypothesised language L that violate C, under a noncumulative hypothesis grammar G.[7] Following Hayes and Wilson (2008), to avoid infinite $E[C]$ without a length limit, the current study limits the maximum string length in L to ℓ_{max}, mirroring the length of the longest string in the input data S. $E[C]$ is then approximated by:

$$E[C] \approx \sum_{\ell=1}^{\ell_{\text{max}}} E_\ell[C] \tag{3}$$

Here, the learner first partitions the input data $S = S_1 \cup S_2 \cup \ldots \cup S_{\ell_{\text{max}}}$ and the hypothesized language $L = L_1 \cup L_2 \cup \ldots \cup L_{\ell_{\text{max}}}\}$ into subsets by string lengths. $E_\ell[C]$ is the expected number of unique strings in each S_ℓ that violate C:

$$E_\ell\,[C] = |S_\ell| \times \textit{Ratio}(C, G, \ell) \tag{4}$$

$\textit{Ratio}(C, G, \ell)$ represents the proportion of strings of ℓ length accepted by G but violating C. This is found by comparing the accepted strings in G and $G' = G \cup \{C\}$, where C is added to G.[8]

$$\textit{Ratio}(C, G, \ell) = \frac{\textit{count}(G, \ell) - \textit{count}(G', \ell)}{\textit{count}(G, \ell)} \tag{5}$$

$\textit{Count}(G, \ell)$ is the count of unique ℓ-length strings in the hypothesis language L accepted by G. Therefore, $\textit{Count}(G, \ell) - \textit{Count}(G', \ell)$ is the number of unique strings that violate C in L.

Table 2 illustrates this calculation with exception-free input data that perfectly align with each hypothesis grammar G. The first row shows an empty hypothesis grammar ($G = \emptyset$) along with input data {CCC, CCV, CVC, CVV, VVV, VCV, VCC, VVC} (where $|S_3| = 8$). $\textit{count}(\emptyset, 3) = 8$, given that the empty hypothesis grammar permits eight potential strings {CCC, CCV, VCC, CVC, CVV, VVV, VCV, VVC} of length 3.

G	Exception-free input data $S_3 = L_3$	E_3[*CC]	E_3[*VV]	E_3[*CV]	E_3[*VC]
$\emptyset$	{CCC, CCV, VCC, CVC, CVV, VVV, VCV, VVC}	3	3	4	4
{*CC}	{CVC, CVV, VVV, VCV, VVC}	0	3	3	3
{*CC, *VV}	{CVC, VCV}	0	0	2	2

Table 2: The list of idealised input data and corresponding hypothesis grammar, as well as expected frequencies for length 3; the input data S_3 here is idealised and identical to the target language L_3

[7] Hayes and Wilson (2008: 427) provides a method to estimate E for cumulative constraints.

[8] Efficient computation can be done using a short-distance algorithm on finite-state automata, such as `shortestdistance` in pynini (Gorman, 2016). The author acknowledges [names omitted for anonymous review] for assistance.

When *CC is added to the intersected grammar, resulting $G' = \{$*CC$\}$, G' only permits five strings {CVC, CVV, VVV, VCV, VVC} (*count*$(\{$*CC$\}, 3) = 5$). The expected frequency of *CC is calculated as follows:

$$\begin{aligned} E\,[\text{*CC}] &= E_3\,[\text{*CC}] \\ &= |S_3| \times \textit{Ratio}(\text{*CC}, \emptyset, 3) \\ &= 8 \times (\frac{\textit{count}(\emptyset, 3) - \textit{count}(\{\text{*CC}\}, 3)}{\textit{count}(\emptyset, 3)}) \\ &= 8 \times (\frac{8-5}{8}) \\ &= 3 \end{aligned} \tag{6}$$

This matches the fact that three strings {CCC, CCV, VCC} violate the potential constraint *CC in the idealised input data L_3 in the first row. Here, $E\,[\text{*CC}] = E_3\,[\text{*CC}]$ because only 3-length strings exist in the input data.

Following this update, ungrammatical strings (violating G) are filtered from the input data S. When G becomes {*CC}, as shown in the second row of Table 2, the input data shrinks to {CVC, CVV, VVV, VCV, VVC} ($|S_3| = 5$). E[*CC] drops to zero, because *CC is already penalised by G (*CC $\in G$). In other potential constraints, for example, $E[\text{*VV}] = |S_3| \cdot (\frac{5-2}{5}) = 5 \cdot \frac{3}{5} = 3$, as three of the five strings allowed by $G = \{$*CC$\}$ violate *VV.

Although alternative calculations such as $O - E$, yielded similar learning results, O/E has the advantage of a clear range from 0 ($O = 0$) to 1 ($O = E$), which is convenient for probing the optimal threshold. During the learning process, a constraint is considered ungrammatical if the O/E ratio falls below a specified threshold ($O/E < \theta$).[9] This comparison is performed at increasing threshold levels, ranging from 0.001 to θ_{max}, also known as *accuracy schedule* (Hayes and Wilson, 2008). This structure prioritises the integration of potential constraints with the lowest O/E values. The maximum threshold θ_{max} demarcates the grammatical and ungrammatical two-factors, and varies accross specific datasets. The current study reports the θ_{max} value that yields the best performance in each dataset. A future direction is to let the learner discover this hyperparameter by maximising the model's performance on the held-out set, assuming a specific noise distribution in the input data. The maximum threshold also accommodates the potential for modelling variability across individuals, as some learners might be more tolerant of lexical exceptions and set a higher maximum threshold.

[9]The current proposal also employs the Normal Approximation technique (Mikheev, 1997; Albright and Hayes, 2003; Hayes and Wilson, 2008) which transforms the O/E ratio into a statistical upper confidence limit. Consequently, with a default confidence level of 0.975, the difference between an O/E of 0/10 and 0/1,000 is not represented as 0 vs 0, but rather as 0.22 vs 0.002, providing more nuanced differentiation. This ensures that potential constraints with a larger discrepancy between O and E values are added to the grammar first, given their potential to better address the variance between observed and expected frequencies. However, implementing this technique did not have a significant impact on learning results in the author's trials.

3.4 Learning Procedure

Building on the concepts above, the Exception-Filtering learner models how a child learner acquire a categorical phonotactic grammmar given the input data. The *learning problem* in the presence of exceptions is formalised as follows: Given the input data S, select a hypothesis grammar G from the hypothesis space, so that G approximates the target grammar $\mathcal{T}$ that defines the target language $\mathcal{L}$.[10] The input data S includes grammatical strings from $\mathcal{L}$ and a limited number of ungrammatical strings outside $\mathcal{L}$, i.e., lexical exceptions, disregarding speech errors, and other noise.

Consider a toy example: given the tier (also the inventory) {C, V}, the target grammar $\mathcal{T} = \{$*CC$\}$. The hypothesis space consists of all possible two-factors on the tier {*CC,*CV,*VV,*VC}. The toy input data $\mathcal{S} = \{$CVC, CVV, VVC, VVV, VCV, CCV$\}$ includes one exception CCV, which violates the target grammar $\mathcal{T}$. Though the toy example limits the string length to three for convenience of discussion, the learner can handle samples with varying word lengths.

[10]The assumption that a single uniform target grammar applies to all native speakers is a simplification. Ideally, the input data should be generated by a single source, such as a parent-teacher. However, in a more realistic learning environment, there might be multiple target grammars across different speakers due to a variety of input data sources, causing variations among native speakers.

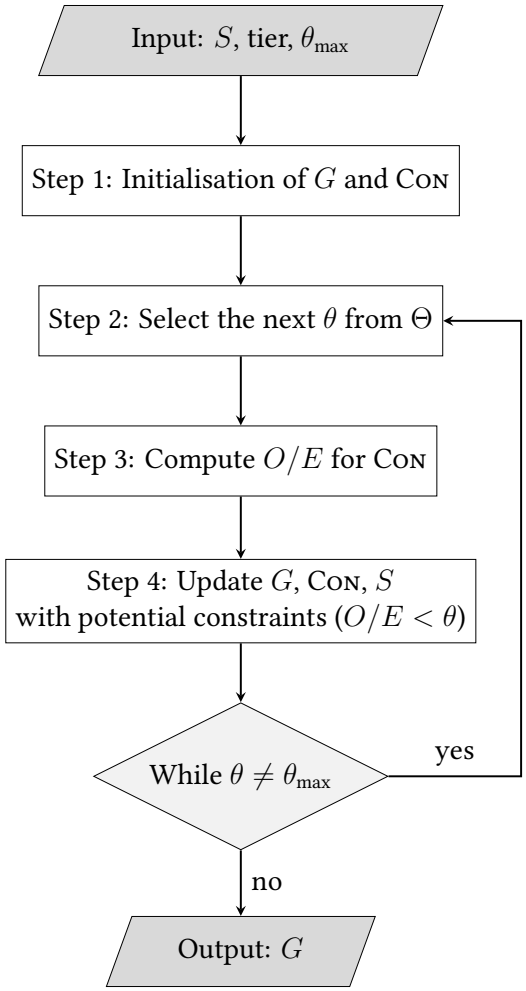


Figure 4: The learning procedure of the Exception-Filtering learner

As visualised in Figure 4,[11] given the input data S, tier, and the maximum O/E threshold θ_{max}, the learner first initialises an empty hypothesis grammar G and hypothesis space Con (Step 1). The learner then selects the next threshold θ from the accuracy schedule Θ (Step 2). Subsequently, the learner computes O/E for each potential constraint within the hypothesis space (Con) (Step 3). Constraints with $O/E < \theta$ are integrated into the G and removed from Con and all lexical exceptions that violate these constraints are filtered out of the input data S (Step 4). This is followed by a reselection of θ, a reevaluation of the values of O/E and an update of G, Con, S (Steps 2, 3 and 4). The learner follows the accuracy schedule and incrementally sets a higher threshold for constraint selection. The iteration continues until the threshold reaches a maximum value ($\theta = \theta_{\text{max}}$), marking the convergence. The following paragraphs illustrate the learning procedure using the toy input data with the exception of *CCV. For convenience of illustration, a simplified accuracy schedule $\Theta = [0.5, 1]$ with a max threshold of 1 is used to avoid too many iterations.

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

3.4.1 Step 1: Initialisation

Given the input data S and tier ({C, V}), the learning process begins with the initialisation of a hypothetical grammar G. Initially, G is an empty set, implying that all possible sequences are assumed to be grammatical prior to the learning procedure. The learner also defines the hypothesis space Con, which encompasses all forbidden two-factors. This initialisation process is shown in Table 3, where the left side shows the initialisation of O and E, and the right side stores the variables:

	O	E	O/E
*VV	0	0	0
*VC	0	0	0
*CV	0	0	0
*CC	0	0	0

G	=	$\emptyset$
Con	=	{*CV, *VV, *VC, *CC}
S	=	{CVC, CVV, VVC, VVV, VCV, CCV}

Table 3: Initialisation

3.4.2 Steps 2 and 3: Select θ, Compute O/E

Following the initialisation, the learner selects the first $\theta = 0.5$ from the accuracy schedule and calculates the observed type frequency O and expected type frequency E for each potential constraint within the hypothesis space Con. In essence, $O[C]$ represents the proportion of strings that violate a potential constraint C in the input data, while $E[C]$ represents the proportion of strings that violate C in the current grammar G.

Consider the toy input data S = {CVC, CVV, VVC, VVV, VCV, CCV} ($|S| = 6$). For the potential constraint *CC, *count*$(G, 3) = 8$ and *count*$(G', 3) = 5$ because three strings violate the updated grammar $G' = \{\text{CC}\}$. The ratio in which a string violates *CC in the expected sample is the same as in the exception-free example above *Ratio*$(\text{*CC}, \emptyset, 3) = 1 - \frac{5}{8} = \frac{3}{8}$. As a result, $E[\text{*CC}] = |S| \cdot$ *Ratio*$(\text{*CC}, \emptyset, 3) = 6 \cdot \frac{3}{8} = 2.25$, as illustrated in Table 4.

	O	E	O/E
*VV	3	2.25	1.33
*VC	3	3	1
*CV	4	3	1.33
*CC	1	2.25	0.44

G	=	$\emptyset$
Con	=	{*CV, *VV, *VC, *CC}
S	=	{CVC, CVV, VVC, VVV, VCV, CCV}
θ	=	0.5

Table 4: Compute O and E

3.4.3 Step 4: Update G, Con, and S (Exception-Filtering)

The learner then stores potential constraints with $O/E < \theta$ in G. Here, the learner updates G with *CC, as shown in Table 5. The sample S is also updated, and strings that contradict the

updated hypothesis grammar are filtered out. In this case, the potential constraint *CC is added to G and removed from Con, and the string CCV is removed from S. This process is depicted in Table 5.

	O	E	O/E
*VV	3	2.25	1.33
*VC	3	3	1
*CV	4	3	1.33
*CC	1	2.25	0.44

G	=	{*CC}
Con	=	{*CV, *VV, *VC, ~~*CC~~}
S	=	{CVC, CVV, VVC, VVV, VCV, ~~CCV~~}
θ	=	0.5

Table 5: Update G, Con, and S

To prevent the overestimation of O/E, the learner filters out ungrammatical strings, including exceptions, from the input data. This is because adding one constraint to the hypothesis grammar has an impact on the expected frequency of other two-factors.[12] For instance, after integrating *CC into the hypothesis grammar, CCV, VCC, and CCC should no longer be considered in the expected frequency count, thereby reducing the expected frequency of *CV and *VC. This mechanism ensures the learner continue the subsequent the learning process without the negative impact of identified lexical exceptions.

3.4.4 Iteration and Convergence

The learner then enters an iterative process and returns to Step 2 to reselect θ and recalculate O and E based on the updated hypothesis grammar G. This iteration is crucial as the values of O and E depend on the current state of G. The process continues until the accuracy schedule is exhausted ($\theta = \theta_{\text{max}}$), indicating that there are no more potential constraints, marking the convergence of learning. In the second iteration of the toy example, after *CC is added to G and removed from Con (hence "NA" in O[*CC] and E[*CC]), θ is reassigned to 1, and no constraint satisfies $O/E < \theta$. $\theta = \theta_{\text{max}} = 1$ indicates the convergence of the learning process. The learnt grammar matches the target grammar $\mathcal{T}$ = {*CC}, as shown in Table 6.

	O	E	O/E
*VV	3	3	1
*VC	3	3	1
*CV	3	3	1
*CC	NA	NA	NA

G	=	{*CC}
Con	=	{*CV, *VV, *VC}
S	=	{CVC, CVV, VVC, VVV, VCV}
θ	=	1

Table 6: Step 2 and 3 after the first iteration

[12] This filtering mechanism does not exist in Hayes and Wilson (2008: 389). Their observed frequency $O[C]$ remains constant throughout the learning process, while $E[C]$ is proportional to the probability of sequences penalised by the constraint C, which is updated by the MaxEnt grammar in each iteration. Technically, this problem is trivial as several hyperparameters can "repair" overestimation and still select correct constraints in their algorithm.

3.5 Summary

To summarise, the Exception-Filtering learner initiates the learning process with an empty hypothesis grammar, allowing all possible sequences. As it accumulates indirect negative evidence from the input data, the learner gradually filters out exceptions, shrinks the space of possible sequences, and updates the hypothesis grammar G with respect to the comparison of the observed and expected type frequency. The learner iteratively filters out lexical exceptions from the input data, rather than accepting them in the hypothesis grammar. The crucial hyperparameter in the learner is the maximum threshold θ_{max}, which depends on the specific datasets.

4 Evaluation

This section aims to provide a clear methodology for evaluating the proposed learning model. Inspired by Hastie et al. (2009), the evaluation in the current study consists of four dimensions (two analytical and two statistical):

1. Scalability: Can the model be applied successfully to a wide range of input data?
2. Interpretability: Can human analysts (linguists) interpret the learnt grammar?
3. Model assessment: Evaluating the performance of the model with new data. This is achieved through the statistical tests against test dataset as discussed below;
4. Model comparison: Comparing the performance of different models.

The current study examines these four dimensions through three case studies in representative datasets: local onsets phonotactics in English and Polish child direct corpora and nonlocal vowel phonotactics in Turkish adult direct corpus. Learning from onset phonotactics helps control the influence of syllable structures and considerably simplifies the learning problem (Daland et al., 2011; Jarosz, 2017; Jarosz and Rysling, 2017). In Turkish, however, learning models are applied to vowel tiers without specified syllabic structures.

Moreover, the current proposal is compared to the learning algorithm proposed by Hayes and Wilson (2008, henceforth HW learner) due to its widespread acceptance in the field and its accessible software (UCLA Phonotactic Learner; https://linguistics.ucla.edu/people/hayes/Phonotactics/), making it an ideal benchmark for comparison. In the case studies, the hyperparameters Max O/E (0.1 to 1) and Max gram size (2 to 3) in the HW learner were fine-tuned so that only the highest performing models across all tests are reported.[13] A 300 Maximum constraint limit was only established in the Turkish case study due to hardware limitations when handling a significantly large corpus. Moreover, the default Gaussian prior is used to reduce overfitting and handle exceptions Hayes and Wilson (2008: 387; $\mu = 0, \sigma = 1$; see more

[13]Similarly, θ_{max} in the Exception-Filtering learner is also reported on the best-performance basis.

discussion on this exception-handling mechanism in §8).[14]

The current study also implements a baseline categorical Tier-based Strictly 2-Local phonotactic learner (henceforth Baseline; capitalised to distinguish from other baseline models), adapted from *memory-seg* learner (Wilson and Gallagher, 2018) and other previous work (Gorman, 2013; Kostyszyn and Heinz, 2022), in which a string is considered grammatical ($g = 1$) if all its two-factors have nonzero frequency in the input data, and ungrammatical ($g = 0$) otherwise.

As the current study proposes a "categorical grammar + exception-filtering mechanism" approach, contrasting it with the HW learner sheds light on the role of categorical grammars, while comparing it with the Baseline learner highlights the significance of the exception filtering mechanism. All models are trained on the same input data.

Although none of the learning models here claim to be the exact algorithm performed by child learners, comparing their learning results and behavioural data provides valuable insights into the underlying mechanisms of phonotactic learning in the face of exceptions. In English and Polish case studies, the learnt grammars are tested on the acceptability judgments from behavioural data. In the Turkish case study, while conducting a new experiment falls outside the scope of the current study, the study approximates the acceptability judgments utilising the experimental data Zimmer (1969). This is in line with the methodology employed by Hayes and Wilson (2008) for deriving acceptability judgments in English from Scholes (1966). Moreover, the learnt grammar is contrasted with the documented grammar as analysed by human linguists. This has been a standard method in phonotactic modelling. For instance, Hayes and Wilson (2008) compared the learnt grammars of Shona and Wargamay with the phonological generalisations in the previous literature. Gouskova and Gallagher (2020) used a method to generate grammaticality labels for nonce words based on phonological generalisations that are experimentally verified (§7).

The major statistical tests for model assessment and comparison are described below:

4.1 Correlation Tests

The correlation between predicted judgments and gradient acceptability judgments, often based on Likert scales, can be assessed using various correlation tests: Pearson's r (Pearson, 1895), Spearman's ρ (Spearman, 1904), Goodman-Kruskal's γ (Goodman and Kruskal, 1954), and Kendall's τ (Kendall, 1938). These values range from -1 (highly negative) to 1 (highly positive).

Pearson's r requires the assumption of linearity, positing that intervals between ratings are of equal size (e.g., the distance between 1 and 2 is the same as between 4 and 5). However, this assumption may not hold for Likert ratings (Gorman, 2013; Dillon and Wagers, 2021), even if they are averaged over participants. Moreover, the Pearson correlation test also requires both variables to be continuous and their relationship to be normally distributed. The categorical grammaticality predicted in the current proposal does not satisfy this requirement. Therefore,

[14]The current study omits the insignificant hyperparameters such as complementation operator, which introduces implicational constraints such as "[s] must precede [+nasal]". This omission has a modest to no impact on learning results, e.g., no difference in the English case and ≈ 0.020 lower Spearman ρ correlation in the Polish case, while this omission ensures a fair and balanced comparison with other models not employing these operators.

Pearson's r is not reported in this study.[15]

Non-parametric tests measuring rank correlations are more appropriate as they make weaker assumptions about the distribution of acceptability judgments (Gorman 2013: 27). Spearman's ρ assumes monotonicity, meaning that the lower values in acceptability consistently correspond to lower levels of predicted grammaticality score. This may also be inaccurate, as subjects may inconsistently assign ratings such as 2, 3, or 4 to intermediate judgments, where a score of 4 could represent less or equal grammaticality as a score of 2.

Hence, the current study introduces two additional measures. In Goodman-Kruskal's γ and Kendall's τ test, pairs of observations (X_i, Y_i) and (X_j, Y_j) from predicted judgments (X) and gradient acceptability judgments (Y) are classified as *concordant*, *discordant*, or *tied*. A pair is considered concordant if the order of elements in X matches that of Y ($X_i < X_j$ implies $Y_i < Y_j$), and discordant if the orders are reversed. If $X_i = X_j$ or $Y_i = Y_j$, the pair is considered a tie.

Goodman-Kruskal's γ calculates the difference between the number of concordant and discordant pairs, normalised by the total number of non-tied pairs: γ = (concordant - discordant) / (concordant + discordant). Tied pairs are ignored in this computation. Kendall's τ penalises tied pairs by modifying the denominator in γ based on the number of tied pairs. Goodman-Kruskal's γ acts as a benchmark when Kendall's τ incurs severe penalty in categorical grammar, which often produces a large number of tied pairs.

4.2 Classification Accuracy

When categorical grammaticality labels are provided in the test data, this paper utilises *binary accuracy* and the F-score as performance measures for predicted grammaticality in the classification task. The binary accuracy represents the proportion of correct predictions of all labels. This value is then separately calculated for "ungrammatical" and "grammatical" labels. F-score is an accuracy metric that takes into account both *precision* and *recall*. Precision is the ratio of true positives to the sum of true positives and false positives. Recall is the ratio of true positives to the sum of true positives and false negatives. The F-score is the harmonic mean of precision and recall $(2 * (\text{precision} * \text{recall})/(\text{precision} + \text{recall}))$, ranging from 0 to 1. A model devoid of false positives obtains a precision score of 1, while one without false negatives achieves a recall of 1. A model without both errors yields an F-score of 1.

To evaluate the HW learner in binary classification, a thresholding method was used to transform the harmony scores of the learnt MaxEnt grammar into categorical grammaticality judgments (Hayes and Wilson, 2008: 385). Specifically, sequences with harmony scores equal to or below a certain threshold were classified as grammatical, whereas those with harmony scores exceeding the threshold were classified as ungrammatical. The optimal threshold was chosen, from the minimum to the maximum of all harmony scores, to maximise the binary accuracy of the learnt MaxEnt grammar. In other words, the current proposal is compared to the maximal performance that a MaxEnt grammar can achieve in binary accuracy.

[15]The author thanks the annonymous reviewers who pointed this out. Consequently, the *temperature* parameter in Hayes and Wilson (2008: 400) is omitted, which only plays a role in their Pearson's correlation test and linear regression.

The following three sections employ the methodologies described above to evaluate the learning results in the case studies of English and Polish onsets and Turkish vowel phonotactics.

5 Case Study: English Onsets

Gorman (2013: 36) has shown that the HW learner does not reliably outperform the baseline learning model based on categorical grammar. This observation was based on the test dataset from studies conducted by Albright (2007); Albright and Hayes (2003) and Scholes (1966). The current study extends this investigation by modelling the learning process from an exceptionful input data set and evaluating the learning results against a novel test dataset drawn from Daland et al. (2011).

5.1 English Input Data

The input of the learner is a "modestly" exceptionful input data, which consists of word-initial clusters taken from 31,985 distinctive word types drawn from the CMU Pronouncing Dictionary. Each of these words has been encountered at least once in the CELEX English database (Baayen et al., 1995; Hayes and Wilson, 2008; Hayes, 2012). This methodology is designed to mirror the learning experiences of children (Pierrehumbert, 2001b).

There are 90 unique onsets in the input data. Table 7 illustrates how the majority of the input data (31,641 to be precise) are classified as nonexotic (7a), while the onsets of 344 words are considered exotic (7b) per Hayes and Wilson (2008). The HW learner yields worse performance when exposed to input data with "exotic" items compared to samples containing only nonexotic items. The current study claims that some, if not all, of these exotic items are lexical exceptions, especially those sequences borrowed from other languages, such as [zl] *zloty* from Polish. Following Hayes and Wilson (2008: 395), [Cj] onsets are removed from the corpus due to considerable phonological evidence indicating that the [j] portion of [Cj] onsets is better parsed as part of the nucleus and rhyme, e.g., *spew* is analysed as [[sp]onset [ju] rhyme][16]. This filtering of [Cj] onsets leads to the input data characterised as "modestly exceptionful" because there are only few remaining exotic onsets.

[16]Gorman (2013: 98) provided a comprehensive review of empirical evidence. For example, [ju] behaves as a unit in language games (Davis and Hammond, 1995; Nevins and Vaux, 2003) and speech errors (Shattuck-Hufnagel 1986: 130).

k	2,764	w	780	s p	313	θ	173	ʃ r	40
r	2,752	n	716	f l	290	s w	153	s p l	27
d	2,526	v	615	k l	285	g l	131	ð	19
s	2,215	g	537	s k	278	h w	111	d w	17
m	1,965	d͡ʒ	524	j	268	s n	109	g w	11
p	1,881	s t	521	f r	254	s k r	93	θ w	4
b	1,544	t r	515	p l	238	z	83	s k l	1
l	1,225	k r	387	b l	213	s m	82		
f	1,222	ʃ	379	s l	213	θ r	73		
h	1,153	g r	331	d r	211	s k w	69		
t	1,146	t͡ʃ	329	k w	201	t w	55		
p r	1,046	b r	319	s t r	183	s p r	51		

(a) Nonexotic input data

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

(b) Exotic input data

Table 7: Type frequency of English onsets in the input data

Several phonotactic patterns are worth noting while interpreting the learnt grammar, especially whether the attested "exotic" onsets such as [sf, zl, zw] are deemed ungrammatical. Moreover, previous studies have emphasised the impact of the Sonority Sequencing Principle (SSP) on English phonotactic judgments. According to the SSP, onsets featuring large sonority rises, such as "stop + liquid" combinations (e.g., [pl, bl, dr]), are generally favoured as being well-formed (Daland et al., 2011).[17] The current study only uses the SSP to better interpret the learnt grammar. Capturing the effects of the SSP on unattested clusters, also known as *sonority projection* (Daland et al., 2011; Jarosz and Rysling, 2017), would require featural representations, which is beyond the scope of this paper.

5.2 Learning Procedure and Learnt Grammar

For the given input data and the tier (all segments of the input data), the Exception-Filtering learner first initialises a hypothesis space for 22 consonants that appear in the input data based on the TSL_2 language, excluding phonemes that never occur at word initial positions such as [x] (as in *loch*) and [ŋ] (*ring*). As a result, the hypothesis space is populated with a total of $22 * 22 = 466$ potential constraints for the English input data. For all case studies, two-factors involving the initial word boundary (#) and each consonant (e.g., *#z) are considered in the hypothesis space, but are ignored in the paper, because they are always deemed grammatical in learnt grammars.

[17] This paper assumes the conventional sonority hierarchy: stops « affricates « fricatives « nasals « liquids « glides (Clements, 1990), and discusses alternative hierarchy from Rubach and Booij (1990) in the Polish case study (§6).

	Stops						Affricates		Fricatives								Nasals		Liquids		Glides	
	p	t	k	b	d	g	t͡ʃ	d͡ʒ	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	l	r	j	w
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
t͡ʃ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d͡ʒ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
θ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
s	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1
ʃ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	1
v	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
z	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
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
l	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	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
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 8: A grammar learnt from the English sample. The first symbol of a two-factor sequence is denoted by the left column, while the second symbols are represented by segments on the penultimate top row. Cells highlighted indicate the attested two-factors within the input data, while blue and red separately correspond to grammatical and ungrammatical two-factors.

The Exception-Filtering learner learns consistent categorical grammars in every simulation, owing to the discrete nature of constraint selection. Arranged according to the sonority hierarchy, Table 8 illustrates the learnt grammar when the maximum threshold $\theta_{\max}$ is set at 0.1, which delivers the optimal performance during the evaluation. The left column denotes the first symbol in a two-factor, while the penultimate top row represents the second symbol. The learner deems grammatical two-factors, such as [pl], as 1, and ungrammatical ones, such as [pt], as 0. The grammatical two-factors such as [bl] in the learnt grammar are all attested, while the attested ungrammatical two-factors such as [pw] indicate detected lexical exceptions. The $\theta_{\max} = 0.1$ demarcates ungrammatical, e.g., [dw] ($O/E = 17/174 \approx 0.098$) and grammatical two-factors, e.g., [ʃr] ($O/E = 40/265 \approx 0.151$).

Interpreting the learnt grammar yields several interesting insights. Only clusters with large

sonority rises are permitted by the learnt grammar, such as "stops + liquids" and "fricatives + liquids", which is consistent with SSP and previous studies (Jarosz 2017: 270), except for [s]-initial two-factors [sp, st, sk]. Moreover, most detected lexical exceptions occur when a consonant is followed by an approximant, as seen in [zl] *zloty*, [sr] *Sri Lanka*, and [pw] *Pueblo*, while these exceptional two-factors all exhibit substantial sonority rises, indicating a conflict between SSP and the learnt grammar.

Furthermore, many learnt segment-based constraints match the MaxEnt grammar learnt in Hayes and Wilson (2008: 397). For instance, the learnt grammar bans sonorants before other onset consonants (*[+sonorant][]; e.g., *rt) and fricative clusters with a preceding consonant (*[][+continuant]; e.g., *sf). Also identified are exceptional two-factors such as *gw, *dw, *θw, also noted by Hayes and Wilson, in which these two-factors are treated as violable constraints instead.

5.3 Model Evaluation in English

This section evaluates whether the learnt grammar approximates the acceptability judgments from the experimental data in Daland et al. (2011). The test dataset includes 96 nonce words of the CC-VCVC structure, e.g., *pr-+-eebid=preebid*. The 48 word-initial CC onsets of these words were randomly concatenated with 6 VCVC tails. There are 18 onsets that never occur as English onsets (unattesteds), e.g., [tl], [rg], and 18 clusters that frequently occur as English onsets (attesteds) as well as 12 clusters that are found only rarely or in loanwords (marginals), e.g., [gw] in *Gwendolyn*, [ʃl] in *schlep* (Daland et al. 2011: 203).

Then each nonce word was rated on a Likert scale, ranging from 1 (unlikely) to 6 (likely), by highly proficient English speakers who were recruited through the Mechanical Turk platform (Daland et al., 2011). Individual scores were not disclosed by the authors, and the test dataset only has averaged Likert ratings over all participants.

Table 9 shows the onsets presented to the subjects and the corresponding type frequency in the input data, the average Likert ratings and the predicted grammaticality (g) of the learnt grammar. Detected exceptions (nonzero frequency but deemed ungrammatical) are highlighted. Notably, the ungrammatical two-factors identified by the Exception-Filtering learner receive low to modest ratings (between 1.325 and 3.124), compared to grammatical two-factors (between 3 and 4.525).

No.	onset	frequency	Likert	g	No.	onset	frequency	Likert	g
1	fr	254	4.525	1	25	dw	17	2.55	0
2	tr	515	4.525	1	26	vr	1	2.5	0
3	gr	331	4.5	1	27	bw	3	2.475	0
4	fl	290	4.1	1	28	θw	4	2.425	0
5	pl	238	4.1	1	29	fw	2	2.4	0
6	ʃr	40	4.025	1	30	pw	1	2.225	0
7	kl	285	4	1	31	zr	0	2.075	0
8	sn	109	3.975	1	32	mr	0	1.85	0
9	pr	1,046	3.95	1	33	tl	0	1.795	0
10	sm	82	3.925	1	34	fn	0	1.7	0
11	kr	387	3.775	1	35	ml	0	1.65	0
12	br	319	3.75	1	36	rl	0	1.625	0
13	dr	211	3.75	1	37	vw	0	1.625	0
14	gl	131	3.725	1	38	dn	0	1.615	0
15	bl	213	3.575	1	39	nl	0	1.6	0
16	tw	55	3.45	1	40	pk	0	1.6	0
17	sw	153	3.2	1	41	km	0	1.575	0
18	ʃl	5	3.125	0	42	rn	0	1.575	0
19	kw	201	3	1	43	rg	0	1.525	0
20	vl	2	3	0	44	lt	0	1.475	0
21	ʃw	3	2.95	0	45	ln	0	1.45	0
22	gw	11	2.675	0	46	dg	0	1.435	0
23	ʃm	5	2.675	0	47	lm	0	1.4	0
24	ʃn	4	2.595	0	48	rd	0	1.325	0

Table 9: Type frequency, averaged Likert ratings, and predicted grammaticality by the learnt grammar of English nonce word onsets; detected exceptions (nonzero frequency and $g = 0$) are highlighted; sorted by averaged Likert ratings

Table 10 provides a performance comparison among the Exception-Filtering ($\theta_{\text{max}} = 0.1$), Baseline, and HW learner (Max O/E = 0.3, Max gram = 2, the same as Hayes and Wilson, 2008). Correlation scores are compared across the entire test dataset as a whole. It should be noted that the test dataset from Daland et al. (2011) excludes several exceptional onsets penalised by the Exception-Filtering learner, such as [*sf].

		Exception-Filtering	Baseline	HW
Correlation (Overall)	Spearman's ρ	0.834	0.839	<u>0.931</u>
	Goodman-Kruskal's γ	0.996	<u>1</u>	0.860
	Kendall's τ	0.690	0.693	<u>0.8</u>

Table 10: Results of the best performance in Exception-Filtering, Baseline, and HW learner; correlation tests are reported with respect to averaged likert ratings in English; best scores are underscored

The reported correlation scores of all models are significantly different from zero at a two-tailed alpha of 0.01. Both the Exception-Filtering and Baseline learners delivered comparable performances[18], while the HW learner demonstrated slightly superior results, especially in terms of Spearman's ρ and Kendall's τ. Interestingly, the close-to-one Goodman and Kruskal's γ observed in both Exception-Filtering and Baseline learners indicates a higher number of tied pairs in nonparametric tests, leading to a marginally reduced Kendall's τ.

Although the Exception-Filtering learner shows a comparable performance on par with other well-established models, it did not stand out in approximating the acceptability judgments of Daland et al. (2011). However, the relatively modest performance of the Exception-Filtering learner in the modestly exceptionful input data sets the stage for improved learning results in the forthcoming sections dealing with highly exceptionful data.

In summary, the proposed learner successfully learns a categorical phonotactic grammar from naturalistic input data of English onsets. The learnt grammar reveals several interesting observations in English phonotactics, and approximates gradient acceptability judgments from the behavioural data in Daland et al. (2011), and managed to deliver a robust performance comparable to benchmark models in a modestly exceptionful input data.

6 Case Study: Polish Onsets

In this section, the Exception-Filtering learner is applied to the input data and gradient behavioural data concerning Polish onsets (Jarosz, 2017; Jarosz and Rysling, 2017).

6.1 Polish Input Data

To model the language acquisition experiences of children, the model was trained on input data that consists of 39,174 word-initial onsets, which is sourced from a phonetically-transcribed Polish lexicon (Jarosz et al., 2017; Jarosz, 2017) derived from a corpus of spontaneous child-directed speech (Haman et al., 2011). There are 384 unique onsets in the input data, and their type frequencies are shown in the appendix B.

[18]The only difference is that Exception-Filtering learner learnt *ʃl which receives an intermediate 3.125 averaged Likert rating, while the Baseline learner deems it grammatical.

	Plosive	Affricate	Fricative	Nasal	Approximant	Trill
Bilabial	p, b			m	w	
Labiodental			f, v			
Alveolar	t, d	t͡s, d͡z	s, z	n	l	r
Alveolo-palatal		t͡ɕ, d͡ʑ	ɕ, ʑ	ɲ		
Retroflex		t͡ʂ, d͡ʐ	ʂ, ʐ			
Palatal					j	
Velar	k, g		x			
Palatalised Velar	kʲ, gʲ					

Table 11: Polish consonant inventory (derived from the input data)

Table 11 shows the consonants that appear in the input data. The current study uses a uniform system for converting orthography to IPA, remaining neutral on the ongoing debate surrounding the specific phonetic properties of certain segments, particularly the retroflex consonants *cz* [t͡ʂ], *drz*/*dż* [d͡ʐ], *sz* [ʂ], and *rz*/*ż* [ʐ] (Jarosz and Rysling, 2017; Kostyszyn and Heinz, 2022). Polish is known for allowing complex onsets (up to four consonants such as [vzdw]) that defy SSP (Jarosz, 2017; Kostyszyn and Heinz, 2022)[19]. For example, a large amount of "glide + stop", "liquid + fricative", "nasal + stop" sequences are attested, such as [wb, rʐ, mkn]. Moreover, many attested onsets are equally or even less acceptable than unattested onsets, as shown in the test dataset below, which provides a unique challenge for the Exception-Filtering learner.

6.2 Learning Procedure and Learnt Grammar in Polish

Similar to the English case study, for the given input data and tier (all segments from the input data), the Exception-Filtering learner initialises possible constraints for 30 consonants that appear in the input data. As a result, the hypothesis space includes a total of $30 * 30 = 900$ two-factors for the Polish input data. As mentioned above, two-factors involving the initial word boundary (#) are ignored because they are all considered grammatical by the learnt grammar. After the learning process, Table 12, arranged according to the sonority hierarchy, illustrates the learnt grammar when $\theta_{\max}$ is set at 0.1, which delivers the optimal performance.

The learnt grammar provides intriguing information on the attested SSP-defying onsets(Jarosz, 2017). Most grammatical two-factors that violate the SSP are obstruent pairs such as "fricative + stop", "fricative + stop", and "fricative + fricative". Rubach and Booij (1990) proposed that stops, affricates, and fricatives have indistinguishable sonority and should be considered as a single category, "obstruents", in the context of the SSP. If one follows this proposition and disregards obstruent initial onsets, most of the remaining SSP-defying two-factors, such as "nasal + obstruent" [rʐ] and "glide + stop" [wd], have relatively low type frequencies and are deemed ungrammatical by the learnt grammar. Only 4 of 900 two-factors are grammatical while defying SSP (have a low or equal sonority rise), namely [lv, rv, mn, mɲ]. In essence, while a comprehensive evaluation

[19]Discussion on the source of Polish SSP-defying phonotactics can be found in Kostyszyn and Heinz (2022, *yer*-deletion) and Zydorowicz and Orzechowska (2017, Net Auditory Distance).

of SSP's role in phonotactic learning is beyond the scope of this study, it is noteworthy that the learnt grammar here shows a viable approach to interpreting SSP-defying onsets in the context of lexical exceptions.

.

	Stop								Affricates						Fricatives									Nasals			Liquids		Glides	
	p	t	k	kʲ	b	d	g	gʲ	t͡s	t͡ɕ	t͡ʂ	d͡z	d͡ʑ	d͡ʐ	f	v	s	z	ɕ	ʑ	ʂ	ʐ	x	m	n	ɲ	l	r	j	w
p	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	1	1	1	1
t	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	1	1	1	1
k	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	1	1	0	1
kʲ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1	1
d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	1	1	1	1	1
g	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	1	1	1	0	1
gʲ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
t͡s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
t͡ɕ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
t͡ʂ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
d͡z	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d͡ʑ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d͡ʐ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	1	1	1	0
v	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	1	1	1	1
s	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	1	1	1	0	1	1	0	1
z	0	0	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	1	1	1	1	1	1
ɕ	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0
ʑ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ʂ	1	1	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1
ʐ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1
x	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	1	0	1
m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ɲ	0	0	0	0	0	0	0	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	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	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	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	0	0	0	0	0	0	0

Table 12: Learnt gammar from Polish input data. The first symbol of a two-factor sequence is denoted by the left column, while the second symbol is represented by segments on the penultimate top row. Cells highlighted indicate the attested two-factors within the input data, while blue and red separately correspond to grammatical and ungrammatical two-factors.

6.3 Model Evaluation in Polish Data

This section evaluates the degree to which the learnt grammar reflects acceptability judgments gathered from experimental data in Polish. The test dataset consists of 159 nonce words, which are constructed from a combination of 53 word-initial onsets (heads) and 3 trisyllabic VCVC(C)V(C) tails. The test dataset also includes 240 attested fillers, varying in word length (1 to 4 syllables) and onset length (0 to 3 consonants). This setting allows for the evaluation of the learner's performance on both attested and unattested sound sequences. Likert ratings were collected from 81 native Polish-speaking adults through an online experiment conducted on Ibex Farm (Jarosz and Rysling, 2017).

No.	onset	frequency	Likert	g	No.	onset	frequency	Likert	g
1	s m	108	4.490	1	28	m ʑ	0	2.881	0
2	g n	7	4.444	1	29	ʑ m	0	2.877	0
3	x r	50	4.420	1	30	f n	0	2.848	0
4	g l	34	4.416	1	31	x ɕ	0	2.802	0
5	ʂ p	53	4.325	1	32	k t͡ʂ	0	2.798	0
6	s n	9	4.259	1	33	ʑ w	0	2.757	0
7	p w	199	4.255	1	34	m d͡ʑ	0	2.745	0
8	ʂ v	0	4.226	0	35	r w	0	2.704	0
9	m r	23	4.193	1	36	r ʐ	5	2.691	0
10	x m	18	4.148	1	37	ɕ x	0	2.568	0
11	p ʂ	1,610	4.078	1	38	d͡ʐ ɲ	0	2.564	0
12	g v	29	4.053	1	39	w ʐ	0	2.556	0
13	t͡ʂ w	12	3.942	1	40	d͡ʐ j	0	2.477	0
14	d ɲ	8	3.757	1	41	l ʑ	0	2.420	0
15	ʐ v	8	3.679	1	42	l j	6	2.412	0
16	g d͡ʑ	10	3.671	1	43	b g	0	2.325	0
17	ʐ m	9	3.642	1	44	w m	0	2.305	0
18	m w	42	3.597	1	45	n w	0	2.284	0
19	ʐ r	1	3.523	0	46	l t͡ʂ	0	2.267	0
20	m n	8	3.453	1	47	ʐ j	2	2.259	0
21	t͡ʂ k	3	3.403	0	48	w r	0	2.160	0
22	t͡ʂ l	1	3.395	0	49	n m	0	2.119	0
23	ʐ w	9	3.144	1	50	n p	0	1.827	0
24	l ɲ	2	3.136	0	51	j d͡ʐ	0	1.687	0
25	m ʐ	1	3.070	0	52	ɲ v	0	1.560	0
26	ʑ l	2	3.004	0	53	j f	0	1.465	0
27	d͡ʐ m	0	2.967	0					

Table 13: Type frequency, averaged Likert ratings, and predicted grammaticality by the learnt grammar of Polish onsets; detected exceptions onsets are highlighted; sorted by Likert

Table 13 shows the onsets presented to the subjects and the corresponding type frequency in the input data, Likert ratings (average by onsets), and the predicted grammaticality (g) of the

learnt grammar. Exceptions detected by the learnt grammar (nonzero frequency and $g = 0$) are highlighted.[20] For instance, [zj] is deemed ungrammatical, which is reflected in its average score of 2.259 on a 1 to 7 Likert scale.

Table 14 shows the correlation with respect to averaged Likert ratings in Table 13. The correlation scores are compared across the entire test dataset as a whole.[21] Correlations in all models significantly differ from zero at a two-tailed alpha of 0.01. In all correlation tests, the Exception-Filtering learner modestly outperforms the Baseline learner. It performs comparably to the benchmark HW learner, with a modestly lower Spearman's ρ and a modestly higher Kendall's τ.

		Exception-Filtering	**Baseline**	**HW**
Correlation (Overall)	Spearman's ρ	0.789	0.712	<u>0.808</u>
	Goodman-Kruskal's γ	<u>0.958</u>	0.823	0.639
	Kendall's τ	<u>0.651</u>	0.586	0.640

Table 14: Results of the best performance in Exception-Filtering, Baseline, and HW learner; correlation tests are approximating averaged Likert ratings in Polish; categorised based on attestedness; best scores are underscored.

The Exception-Filtering learner identified more exceptional two-factors within the Polish input data. Moreover, its performance relative to the benchmark models improved compared to the English case study and surpassed the Baseline learner that lacks the exception-filtering mechanism. These findings highlight the value of the exception-filtering mechanism in phonotactic learning, particularly when dealing with exceptionful real-world corpora.

To summarize, the Exception-Filtering learner, trained on Polish child-directed corpus, has illustrated its potential in extracting categorical grammars that approximate acceptability judgments. The performance of the model is on par with the HW learner in Spearman's ρ, and is modestly outperforms the benchmark HW learner and the Baseline learner in both Goodman-Kruskal's γ and Kendall's τ test, demonstrating its capability in approximating acceptability judgments. These results further substantiate the potential of the Exception-Filtering learner in inducing phonotactic patterns from realistic corpora.

[20] There is a substantial variability among participants in the use of the Likert scale. Some participants tend to assign higher average Likert ratings (up to 6.006), while others lean toward lower average Likert ratings (down to 1.748). The standard deviation of Likert ratings for each word spans a wide range from 0 to 2.88, demonstrating the variability in participants' responses.

[21] The correlation scores are not reported separately for attested (type frequency > 0) and unattested (type frequency $= 0$) sequences as in Jarosz and Rysling (2017) because the Exception-Filtering learner uniformly assigns them a score of 0 to unattested sequences, resulting in a standard deviation of zero and nullify the correlation tests.

7 Case Study: Turkish Vowel Phonotactics

This section tests the Exception-Filtering learner's capability in capturing nonlocal vowel phonotactics from highly exceptionful input data drawn from an adult-directed corpus in Turkish.

7.1 Turkish Vowel Phonotactics

This section applies the current proposal to vowel phonotactic patterns in Turkish. Turkish vowels are shown in Table 15. Turkish orthography is converted to IPA, including *ö* [ø], *ü* [y], and *ı* [ɯ].

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

Table 15: Turkish vowel system

Turkish vowel phonotactic patterns are summarised as follows, adapted from Kabak (2011):

1. **Backness harmony**: All vowels must agree in terms of frontness or backness.
2. **Roundedness harmony**: High vowels must also agree in roundness with the immediately preceding vowel; hence, no high-rounded vowels can be found after the unrounded vowels within a word.
3. **No non-initial mid round vowels**: No mid round vowels (i.e. [o] and [ø]) may be present in a noninitial syllable of a word, which means that they cannot follow other vowels.

First, a vowel cannot follow another vowel with a different [back] value ("backness harmony"). This is clearly demonstrated in morphophonological alternations, as shown in Table 16 (a) and (b), adapted from Gorman (2013: 46). For instance, when a plural suffix is added to the root /pul/ "stample", [lɑr] instead of [ler] surfaces "stamples". This can be attributed to the phonotactic constraint that restricts the nonlocal u…e co-occurrence. In contrast, when /køy/ village" is combined with /lAr/, the resulting term is [køyler] "villages", demonstrating the nonlocal *ø…ɑ co-occurrence restriction. However, exceptions against this generalisation exist both within roots and across root-affix boundaries, as illustrated in examples (c) and (d) in Table 16. For example, both the root [silɑh] "weapon" and the derived form [silɑh-lɑr] "weapons" violate the restrictions of vowel co-occurrence *i…ɑ.

	NOM.SG.	NOM.PL.	meaning	
a.	ip	ip-ler	“rope”	(Clements et al., 1982)
	køy	køy-ler	“village”	
	yyz	yyz-ler	“face”	
	kɯz	kɯz-lɑr	“girl”	
	pul	pul-lɑr	“stamp”	
b.	neden	neden-ler	“reason”	(Inkelas et al., 2000)
	kiler	kiler-ler	“pantry”	
	pelyr	pelyr-ler	“onionskin”	
	boğɑz	boğɑz-lɑr	“throat”	
	sɑpɯk	sɑpɯk-lɑr	“pervert”	
c.	mezɑr	mezɑr-lɑr	‘grɑve”	(Inkelas et al., 2000)
	model	model-ler	“model”	
	silɑh	silɑh-lɑr	“weapon”	
	memur	memur-lɑr	“official”	
	sɑbun	sɑbun–lɑr	“soap”	
d.	etol	etol-ler	“fur stole”	(Göksel and Kerslake, 2004)
	sɑɑt	sɑɑt-ler	“hour, clock”	
	kɑhɑbɑt	kɑhɑbɑt-ler	“fault”	

Table 16: Turkish nominatives that undergo backness harmony (a, b) and exceptions (c, d)

In the second phonotactic constraint related to roundness harmony, a high vowel cannot follow another vowel with a different [round] value (“roundness harmony”), as shown in Table 17 (a). Table 17 provides examples of this pattern. Yet again, exceptions are noted, such as in the root [boğɑz] “throat” and its derived forms.[22]

Last but not the least, mid round vowels [ø] and [o] are typically restricted to initial positions in native Turkish words. This is evident in words like [ødev] “homework” and *oyun* “game”. Consequently, these vowels should not follow any other vowels, for example, *ɑ…ø and *e…o. However, in loanwords, mid round vowels may occur freely in any position.

Generally, a substantial number of exceptions to these phonotactic patterns arise from compounds and loanwords (Lewis, 2001; Göksel and Kerslake, 2004; Kabak, 2011). For example, the compound word [bugɯn] “today” ([bu] “this” + [gɯn] “day”) violates the roundness harmony; the loanword [piskopos] borrowed from Greek *epískopos* “bishop” violates both the roundness harmony and the constraint on non-initial mid round vowels.

Despite many exceptions, these generalisations are not only well-documented in the literature, including Underhill (1976: 25), Lewis (2001: 16), Göksel and Kerslake (2004: 11), and Kabak

[22] A unique case of exceptions is caused by the phenomenon of root-internal *labial attraction*, where ɑC_[+labial]u is produced given the intervocalic labial consonant, as seen in [sɑbur] “patient” (Lees, 1966). However, this pattern is not internalised by all native speakers, as shown in the ratings of nonce words by native speakers (Zimmer, 1969). Modelling labial attraction would require extending the tier from vowel to labial consonants. This task falls beyond the scope of the current study, which treats these cases as exceptions to roundness harmony, leaving the detailed investigation of labial attraction for future research.

	NOM.SG.	DAT.SG.	GEN.SG.	meaning	
a.	ip	ip-i	ip-in	“rope”	(Clements et al., 1982)
	kɯz	kɯz-ɯ	kɯz-ɯn	“girl”	
	sɑp	sɑp-ɯ	sɑp-ɯn	“stɑlk”	
	køy	køy-y	køy-yn	“villɑge”	
	son	son-u	son-un	“end”	
b.	boğɑz	boğɑz-ɯ	boğɑz-ɯn	“throat”	(Inkelas et al., 2000)
	pelyr	pelyr-y	pelyr-yn	“onionskin”	
	døviz	døviz-i	døviz-in	“currency”	
	yɑmuk	yɑmuğ-u	yɑmuğ-un	“trapezoid”	
	ymit	ymit-i	ymit-in	“hope”	

Table 17: Turkish round harmony patterns in morphophonological alternations (a) and exceptions (b) (Gorman, 2013)

(2011: 4), but also supported by experimental studies (Zimmer, 1969; Arik, 2015). Furthermore, recent acquisition studies reveal that some harmony patterns are discernible by infants as early as six months old, who extract and pay attention to the harmonic patterns present in their language environment, filtering out any disharmonic tokens (Altan et al., 2016).

Another layer of complexity in Turkish vowel phonotactics comes from root harmony. Turkish vowel phonotactic constraints are applicable within roots and across morpheme boundaries (Zimmer, 1969; Arik, 2015), while it is still a matter of debate whether harmony patterns in the domain of roots should be analysed as active phonological processes given the existence of exceptions in disharmonic roots (Kabak, 2011: 17), some of which may originate from loanwords. However, from the perspective of phonological learning, these roots constitute a significant part of the input data exposed to human learners, as most Turkish roots can stand alone.

Therefore, Turkish vowel phonotactic patterns pose a unique challenge for phonological learning: How does the learner acquire vowel phonotactic generalisations from both roots and derived forms, despite the high level of lexical exceptions within the input data?

7.2 Turkish Input Data and Learning Procedure

The current study uses the Turkish Electronic Living Lexicon (TELL; http://linguistics.berkeley.edu/TELL/ Inkelas et al., 2000) as input data, which consists of ≈ 66000 roots and the elicited derived forms (root + affixes) produced by two adult native English speakers.[23] Table 18 shows the type frequency of all nonlocal two-factors on the vowel tier in TELL. Two-factors that follow the Turkish vowel phonotactics introduced above are highlighted. This adult-direct corpus is a great testing ground for evaluating the role of the exception-filtering mechanism. Notably, every nonlocal two-factor has a nonzero frequency in this dataset. Therefore, any phono-

[23]During the learning process, morpheme boundaries are disregarded on the vowel tier. The current study acknowledges the presence of derived forms in the input data, but remains neutral on whether these forms are stored as whole words within the lexicon (see discussion on whole-word storage in Lignos and Gorman, 2012).

tactic learner that assumes every attested two-factor to be grammatical would invariably conclude that all combinations are allowed and completely miss the vowel harmony patterns.

$\sigma_1 \downarrow \sigma_2 \rightarrow$	i	e	y	ø	ɯ	ɑ	u	o
i	10,950	4,768	221	123	768	3216	202	1,000
e	15,984	7,130	591	129	663	2873	625	760
y	422	2,944	2,465	43	121	750	177	59
ø	32	982	1,179	27	19	98	18	19
ɯ	247	392	17	60	6,360	3,009	93	207
ɑ	4,369	3,197	394	308	16,887	10267	1,526	1,656
u	475	606	147	40	153	3035	4,058	155
o	857	787	139	42	99	2,591	3,737	684

Table 18: The type frequency of two-factors in the input data; cells of documented grammatical two-factors are highlighted.

Similar to previous case studies, for the given input data and tier (all vowels from the input data), the Exception-Filtering learner initialises possible constraints for 8 Turkish vowels, which yields 64 two-factors in the hypothesis space. The optimal maximum O/E threshold is 0.5. The learnt grammar is illustrated in the first test dataset below.

7.3 Model Evaluation

This section evaluates the learning models in two separate test datasets below.

7.3.1 The First Test Dataset (Categorical Labels)

The first test dataset consists of 64 nonce words in the template of [tV_1kV_2z], such as [tokuz], representing all possible two-factors on the vowel tier. Each word is categorically labelled 1 ("grammatical"; 16 in total) or 0 ("ungrammatical": 48 in total) based on the aforementioned well-documented phonotactic generalisations.[24] Only roots are included in this analysis, as the learning model disregards morpheme boundaries.

It is important to note that individual variability is expected and that the grammaticality labels here may not match the exact target grammar of *every* native speakers. However, these categorical labels are supported by Zimmer (1969)'s behavioural experiment (see appendix A for details); the majority of participants preferred the harmonic to disharmonic roots in a yes/no rating task, which provides the evidence for the psychological reality of Turkish vowel phonotactic patterns encoded in the first test dataset. In other words, the first test dataset aims to evaluate how well the learnt grammar mirrors the categorical phonotactic judgments of the *majority* of participants in Zimmer (1969)'s experiment. This follows the common practise in previous

[24]This approach avoids any sampling bias that might arise from manually reducing or increasing the amount of either categories.

computational studies when acceptability judgments of nonce words in the test dataset are not accessible. For instance, Gouskova and Gallagher (2020) manually labelled the categorical grammaticality of nonce words in the test dataset based on documented phonotactic generalisations supported by behavioural experiments (Gallagher, 2014, 2015, 2016).

Table 19 summarises the tests of classification accuracy on the first test dataset with categorical labels. The Baseline learner miscategorised all nonce words as grammatical, which caused the Baseline learner to achieve perfect recall but at the expense of the lowest precision (0.238), F-score (0.385), and binary accuracy (0.238) due to false positives.

		Exception-Filtering	**Baseline**	**HW**
Classification accuracy	overall	<u>0.969</u>	0.238	0.906
	ungrammatical	<u>1</u>	0	0.875
	grammatical	0.875	<u>1</u>	0.917
	F-score	<u>0.933</u>	0.385	0.824
	precision	<u>1</u>	0.238	0.778
	recall	0.875	<u>1</u>	0.875

Table 19: Performance comparison of Exception-Filtering, Baseline, and HW learner in the first test dataset (categorical labels); best scores are underscored

As discussed in §4, the harmony scores of the benchmark HW learner are transformed into categorical labels to produce its highest binary accuracy. However, even at its best performance (Max $O/E = 0.7$, $n = 3$, vowel tier: [high], [round], [back], [word boundary]), the HW learner displayed higher error rates in the classification of Turkish phonotactics than the Exception-Filtering learner.

$\sigma_1 \downarrow \sigma_2 \rightarrow$	i	e	y	ø	ɯ	ɑ	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	0	0	0	0	0	0	0
ɯ	0	0	0	0	1	1	0	0
ɑ	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

(a) Exception-Filtering

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

(b) HW

Figure 5: Compare the learnt grammars of (a) Exception-Filtering learner and (b) HW learner

When tested against these categorical labels, the Exception-Filtering learner ($\theta_{\max} = 0.5$) demonstrated outstanding performance in binary classification with an F-score of 0.933, and a

total binary accuracy of 0.969. Figure 5 shows the comparison between the grammars acquired by the Exception-Filtering learner (a) and the benchmark HW learner (b). A score of 0 indicates that a two-factor has been classified as ungrammatical, whereas a score of 1 designates it as grammatical. In (b), the degress of shading is proportional to the negative harmony scores, which is rescaled according to the minimum and maximum harmony score.

Compared to phonotactic generalisations in Turkish, the learnt grammar in the Exception-Filtering learner predicts two false negatives, which are reflected in the relatively lower recall (0.875) in classification accuracy. These two mismatches have an unexpectedly low type frequency (ø…e: 982; ø…y: 1,179), compared to other grammatical two-factors. On the contrary, the errors of the learnt MaxEnt grammar are mostly false positives misled by their high type frequency, such as e…ɑ (2,873), ɑ…i (4,369), ɑ…u (1,526), and ɑ…e (3,197). The Exception-Filtering learner avoids these false positives by categorically penalising these exceptional two-factors.[25]

7.3.2 The Second Test Dataset (Approximated Acceptability Judgments)

The purpose of the second test dataset is to demonstrate that the learnt categorical grammar can approximate the acceptability judgments in the behavioural data. The second testing data includes 36 nonce words in Zimmer (1969), and takes the proportion of "yes" responses averaged across participants to approximate the acceptability judgments of native speakers. The data show a gradient transition from harmonic, e.g., [temez] receives $19/23 \approx 0.826$ to disharmonic words e.g., [temɑz] $3/23 \approx 0.130$. This method is similar to Hayes and Wilson (2008)'s approach to create gradient acceptability judgments from the Scholes (1966) experiment, following previous studies (Pierrehumbert, 1994; Coleman and Pierrehumbert, 1997).

Table 20 presents the results of the statistical tests. The Baseline learner is omitted due to its lack of standard deviation, which makes correlation tests inapplicable. Notably, while correlations in all models differ significantly from zero at a two-tailed alpha of 0.01, the Exception-Filtering learner scored higher than the benchmark HW learner in all tests.

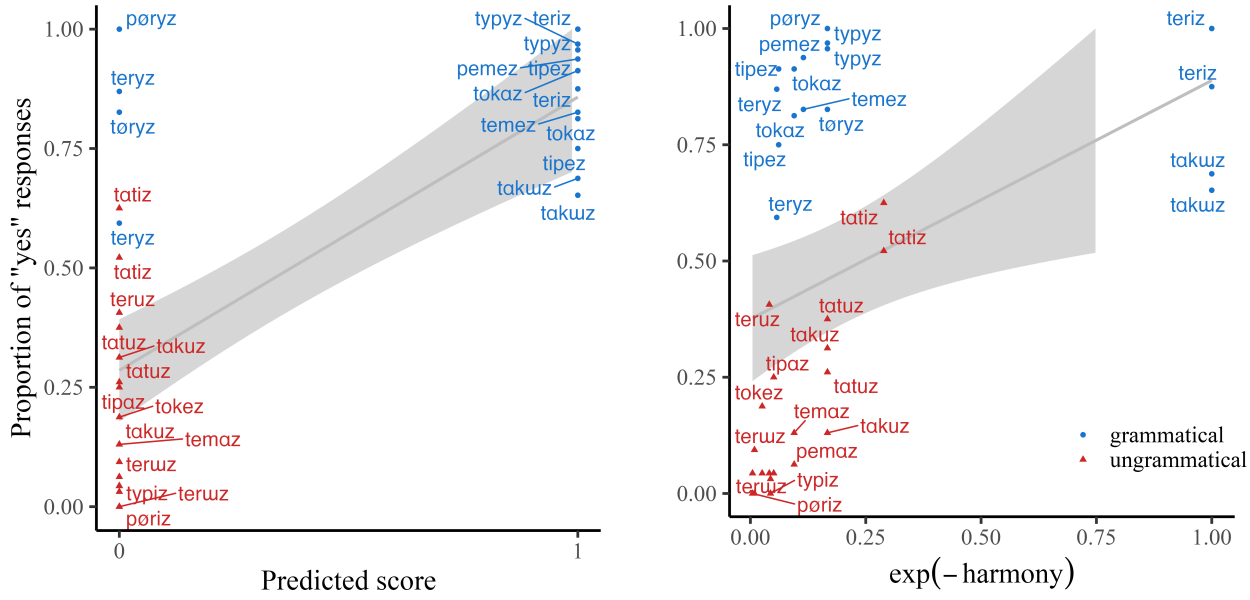
		Exception-Filtering	**HW**
Correlation tests	Spearman's ρ	<u>0.699</u>	0.651
	Goodman-Kruskal's γ	<u>0.860</u>	0.527
	Kendall's τ	<u>0.584</u>	0.500

Table 20: Performance comparison of Exception-Filtering and HW learner in the second test dataset adapteds from Zimmer (1969)'s experiment; best scores are underscored.

Figure 6 visualises the distribution of predicted score against the approximated acceptability in both Exception-Filtering and HW learner. A simple linear regression line is fitted here, where the predictor (x-axis) is the predicted grammaticality score in the Exception-Filtering learner, and

[25]The current study also tests the case when the Exception-Filtering learner does not filter out the identified lexical exceptions from the input data, it falsely classifies two more cases as ungrammatical: ɯ…ɑ (frequency 3,009) and u…ɑ (frequency 3,035).

the exponentiated negative harmony score in the HW learner. The outcome (y-axis) is the proportion of "yes" responses in Zimmer (1969), which approximates the acceptability judgments. The predicted scores of the Exception-Filtering learner cluster at 0 and 1, while the exp(−harmony) is on a continuum.[26]



(a) Exception-Filtering

(b) HW

Figure 6: Regression models based on the learning results of two learners; expected grammaticality is hightlighted based on the documented phonotactic generalisations; overlapped words are omitted on the plots.

Both regression models reject the null hypothesis that the predicted judgments have no effect on the proportion of "yes" responses (Exception-Filtering: residual deviance $= 2.264, p < 0.001$; HW: residual deviance $= 4.073, p = 0.013$), at an alpha level of 0.05. Furthermore, Figure 6 shows that the Exception-Filtering learner is capable of categorically penalising lexical exceptions that are underpenalised by the HW learner, such as ɑ...i in [tɑtiz].

To summarise, the Exception-Filtering learner trained using a large-scale Turkish corpus acquired the documented vowel phonotactics in Turkish except for two mismatches. The Exception-Filtering learner not only succeeded in classifying grammatical and ungrammatical words, but also achieved a high correlation between the predicted judgment and the approximated acceptability judgment of nonce words from previous behavioural experiment. This result indicates the capability of the Exception-Filtering model in modelling phonotactic patterns with exceptions.

[26] As harmony scores range from 0 to positive infinity, the corresponding values of exp(−harmony) decrease from 1 to 0, approaching but never reaching 0 as harmony scores approach infinity. Therefore, the range of exp(−harmony) for harmony in $[0, +\text{inf})$ is $(0, 1]$. This value should not be mistaken as probability, despite their similar ranges.

8 Discussion

To summarise the case studies, in terms of scalability and interpretability, the categorical grammars learnt in the case studies of English and Polish onset phonotactics largely align with the Sonority Sequencing Principle that penalises most sequences with low sonority rises. The proposed learner also successfully generalised Turkish vowel phonotactics from highly exceptionful input data with both roots and derived forms. When it comes to model assessment and comparison, the grammaticality scores generated by the learnt grammars closely approximate the acceptability judgments observed in behavioural experiments and demonstrate competitive performance in model comparisons, highlighting the effectiveness of the exception-filtering mechanism. The following section discusses topics that arise from the current study and outlines directions for future work.

8.1 Extragrammatical Factors

As elaborated in §2, this research adopts the competence-performance dichotomy within the dual-route model (Pinker and Prince, 1988; Zuraw, 2000; Zuraw et al., 2021). Within this framework, extragrammatical factors are conceptualised as originating from two main sources: performance-related variables and lexicon-related variables. Performance-related variables include task effects (Armstrong et al., 1983; Gorman, 2013), individual differences, auditory illusions (Kahng and Durvasula, 2023), etc. Lexicon-related variables include lexical information such as lexical similarity (Bailey and Hahn, 2001, 2005; Avcu et al., 2023) and frequency (Frisch et al., 2000; Ernestus and Baayen, 2003).

In the current study, extragrammatical factors serve dual roles in both the learning and evaluation phases of phonotactic learning. In the learning phase, the learner can use extragrammatical information, such as the lexical frequency. For instance, the proposed exception-filtering mechanism in this study distinguishes between grammatical sequences and lexical exceptions by using type frequency.

In the evaluation phase, in tandem with the learnt grammar, extragrammatical factors contribute to acceptability judgments in behavioural experiments. For example, previous studies have shown that lexical similarity and frequency are significant predictors of acceptability judgments (Bailey and Hahn, 2001, 2005; Frisch et al., 2000). Therefore, a comprehensive evaluation of a learnt grammar against acceptability judgments should take these factors into account. In future research, this evaluation could be conducted by adopting a mixed-effects regression model, in which the grammaticality score is treated as a fixed effect and extragrammatical factors are treated as other effects. Although this approach restricts the explanatory power of the grammar, it also factors out the sources of acceptability judgments. It is worth noting that such a complex model would require an extensive amount of data.

.

8.2 Accidental Gaps

Accidental gaps, the unattested but grammatical sequences emerging from the lexicon-grammar discrepancy, pose a significant challenge to phonotactic learning. Given that there are logically infinite numbers of grammatical strings and only some of them are associated with lexical meaning, gaps in the input data are inevitable. These accidental gaps can lower the O/E ratio because expected sequences are absent in the input data, which could potentially lead the learner to misinterpret these sequences as ungrammatical. This issue does not cause severe problems in the current proposal because the learner can potentially avoid the misgeneralisation of accidental gaps by adjusting the maximum threshold. However, this is not a fundamental solution and places an excessive burden on a simple statistical criterion.

A more principled solution to the challenge of accidental gaps is to incorporate feature-based constraints, as suggested by Wilson and Gallagher (2018). Underrepresented segmental two-factors in the input data may exhibit high frequency when analysed from a feature-based perspective. For instance, in English, b[+approximant] is highly frequent (e.g., br, bl), except for [bw] which only has three unique occurrences. In contrast, all segmental two-factors are unattested for b[-approximant] (e.g., bn, bg, bt). A feature-based grammar can penalise [-approximant] after b, but allow b[+approximant], hence avoiding overpenalising accidental gaps with [bw] onsets. By considering the entire natural class, the grammar can recognise subsegmental patterns that are overlooked in segmental representation.

It is feasible to integrate featural representations into the current approach using the generality heuristics in Hayes and Wilson (2008) and the bottom-up strategies proposed by Rawski (2021). The current study offers a straightforward demonstration of the concept here: consider a simplified feature system illustrated in Table 21, a feature-based Exception-Filtering learner initialises the most general feature-based potential constraints, e.g.,[+F][+F], *[+F][+G], etc.

	F	G
C	-	+
V	+	-

Table 21: Simplified feature system

After selecting the next threshold from the accuracy schedule, and computing the O/E for each possible two-factor, the learner adds a two-factor to the hypothesis grammar if (1) the two-factor is not implied by any previously learnt constraints, and (2) the O/E of the two-factor is lower than the current threshold. For example, a constraint such as *[+G][+G] would imply more specific two-factors like *[+G][+F,+G], *[+F,+G][+G], but not *[+F][+F]. Therefore, if *[+G][+G] is already learnt, the learner will not consider the implied *[+G][+F,+G] regardless of its O/E value. The learning process continues until all thresholds have been exhausted. The next step of the current study is to incorporate more learning strategies proposed in Hayes and Wilson (2008) and Rawski (2021) to optimise the learner for natural language corpora.

8.3 Hayes & Wilson (2008) Learner

The Exception-Filtering learner drew inspiration from probabilistic approaches, espeically the benchmark HW learner, which learns a Maximum Entropy Grammar (Goldwater and Johnson, 2003; Berger et al., 1996) from input data. The HW learner adjusts constraint weights to maximise the likelihood of the observed data predicted by the hypothesis grammar, also known as Maximum Likelihood Estimation (MLE), aiming to approximate the underlying target grammar. However, MLE operates under the assumption that the observed input data consist only of grammatical words. This paper has shown that such an assumption may falter in the presence of lexical exceptions.

Interestingly, although the HW learner also uses the O/E criterion in constraint selection, it cannot exclude lexical exceptions from the input data even with the correct constraints selected. The principle of MLE prevents the probabilistic grammar from assigning a zero probability to observed lexical exceptions and from completely excluding these anomalies. The underpenalisation of lexical exceptions can compromise generalisations for nonexceptional candidates (Moore-Cantwell and Pater, 2016). For example, in the Turkish case study, the HW learner underpenalised the highly frequent disharmonic patterns such as ɑ…i in [tɑtiz] (Figure 6).

This issue has motivated several interesting proposals to handle exceptions within the HW learner. Hayes and Wilson (2008: 386) added a Gaussian prior to prevent overfitting by adjusting the standard deviation σ of the Gaussian distribution for constraint weights. Its effectiveness depends on the distribution of noise in specific input data. Another strategy is to include lexically specific constraints in the hypothesis space to handle lexical exceptions (Pater, 2000; Linzen et al., 2013; Moore-Cantwell and Pater, 2016; Hughto et al., 2019; O'Hara, 2020). Lexically specific constraints such as *sf$_i$ would normally penalise the sequence [sf], except when it is in the indexed lexical exception *sphere*$_i$. In this way, the learnt grammar is able to allow exceptions without compromising the generalisations for nonexceptional candidates. Meanwhile, nonce words are evaluated under the general constraints of the grammar, as they would never violate any established lexically indexed constraints. However, lexically specific constraints considerably escalate the computational complexity of the learning model due to the exponential growth of the hypothesis space with respect to the size of input data. Therefore, lexically specific constraints seem more likely to emerge after phonotactic learning when a grammar has been established and lexical exceptions have been identified.

Both proposals above handle the exception-related overfitting problem through the incorporation of a regularizer function during maximum likelihood estimation. An open question is whether the HW learner can be improved by incorporating the exception-filtering mechanism advocated in the current proposal, so that identified anomolies can be removed from input data during the learning process.

8.4 *O/E* and Alternative Criteria

Both the Exception-Filtering learner and the HW learner employ a "greedy" algorithm that selects constraints whenever O/E is below the selected threshold. This approach, while computationally efficient, does not guarantee the discovery of a globally optimal grammar, given that the addition

of one constraint may influence the O/E of others.[27] As the learning model does not possess the capacity to "look ahead", it becomes vital for the analyst to thoroughly examine the learning results across various threshold levels to uncover potential implications and enhancements. In the context of learning phonotactic grammars from exceptionful data, the O/E criterion has proven to be an effective measure in case studies.

An alternative strategy, such as the use of a depth-first search algorithm, could circumvent local optima by allowing the learner to examine future constraints before committing to the current one. However, this method comes with a considerable increase in computational complexity.

To ultimately solve the problem of local optima, a future direction is to consider other criteria, such as the *gain* s as per Della Pietra et al. (1997); Berent et al. (2012), and the Tolerance Principle as per Yang (2016).

The gain criterion was originally designed for well-defined probabilistic distributions, and its convex property ensures that the added constraints approximate a global optimum. Generalising this criterion to the current proposal based on categorical grammars involves some nontrivial adjustments, especially deriving a probabilistic distribution from categorical grammars.

The Tolerance Principle proposes that a rule will generalise if the number of exceptions does not surpass the number of words in the category N divided by the natural log of N ($N/\ln N$). This concept bears close resemblance to the current proposal, where a constraint is added to the grammar if the observed exceptions do not exceed a specified threshold with respect to the expected type frequency. While this constitutes a promising avenue for future research, it is worth noting that the Tolerance Principle was not originally formulated with phonotactic learning in mind, and it requires adjustment in defining the scope of constraint application.

To summarise, while the current constraint selection criterion provides the best balance between mathematical rigour and performance, other criteria could be incorporated into the "categorical grammar + exception-filtering mechanism" approach in the future.

8.5 Other Future Directions

The current study represents an initial step towards understanding the interplay between lexical exceptions and phonotactic learning. The primary objective of this study has been to address the issue of exceptions, rather than developing an all-encompassing learning model. This has led to significant simplifications in the proposed learning model. Therefore, the natural next step is to enhance the current proposal towards a more comprehensive model.

First, this study uses a simplified noncumulative categorical grammar, while experimental evidence has indicated a cumulative effect on phonotactic learning (Breiss, 2020; Kawahara and Breiss, 2021). A promising future direction involves adapting the current proposal to accommodate a cumulative grammar, which would subsequently alter the assignment of grammaticality and the calculation of O/E.

Secondly, while this study abstracts away from a comprehensive evaluation of the role of SSP in phonotactic learning, the learnt grammar in Polish shows a viable approach to interpret SSP-defying onsets in the context of lexical exceptions (Jarosz, 2017).

[27] The author thanks the reviewer for pointing out this issue.

Moreover, this study prespecifies tiers for the hypothesis space. In the future, it would be beneficial to integrate an automatic tier induction algorithm based on the principles proposed in previous studies (Jardine and Heinz, 2016; Gouskova and Gallagher, 2020).

Finally, the current study does not model phonological alternations. This problem involves inference of the underlying representations, which has been studied in idealised input data (Hua et al., 2021; Hua and Jardine, 2021). A potential future direction involves the simultaneous inference of a phonotactic grammar, underlying representations, and phonological processes from realistic corpora, while filtering out lexical exceptions.

9 Conclusion

In conclusion, this research represents a significant step forward in two key areas. First, it pioneers a "categorical grammar + exception-filtering mechanism" approach for learning categorical grammars from naturalistic input data with lexical exceptions. Furthermore, while the current study primarily focuses on the learning of categorical grammars, it lays the groundwork for integrating learnt grammars with extragrammatical factors to model behavioural data. Therefore, this study marks initial steps in reevaluating the capacity of categorical grammars to approximate human judgments.

References

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

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.

Algeo, J. (1978). What consonant clusters are possible? *Word*, 29(3):206–224.

Altan, A., Kaya, U., and Hohenberger, A. (2016). Sensitivity of turkish infants to vowel harmony in stem-suffix sequences: preference shift from familiarity to novelty. In *Proceedings of the 40th Boston University Conference on Language Development*.

Archer, S. L. and Curtin, S. (2016). Nine-month-olds use frequency of onset clusters to segment novel words. *Journal of experimental child psychology*, 148:131–141.

Arik, E. (2015). An experimental study of turkish vowel harmony. *Poznan Studies in Contemporary Linguistics*, 51(3):359–374.

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

Avcu, E., Newman, O., Ahlfors, S. P., and Gow Jr, D. W. (2023). Neural evidence suggests phonological acceptability judgments reflect similarity, not constraint evaluation. *Cognition*, 230:105322.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The celex lexical database (release 2). *Distributed by the linguistic data consortium, University of Pennsylvania*.

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

Bailey, T. M. and Hahn, U. (2005). Phoneme similarity and confusability. *Journal of memory and language*, 52(3):339–362.

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.

Bloomfield, L. (1933). *Language*. Holt.

Breiss, C. (2020). Constraint cumulativity in phonotactics: evidence from artificial grammar learning studies. *Phonology*, 37(4):551–576.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.

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

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper & Row, New York.

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. (1990). The role of the sonority cycle in core syllabification. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 283–333. Cambridge University Press.

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

Coleman, J. and Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Dai, H., Mayer, C., and Futrell, R. (2023). Rethinking representations: A log-bilinear model of phonotactics. *Proceedings of the Society for Computation in Linguistics*, 6(1):259–268.

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

Davis, S. and Hammond, M. (1995). On the status of onglides in american english. *Phonology*, 12(2):159–182.

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.

Dillon, B. and Wagers, M. W. (2021). *Approaching Gradience in Acceptability with the Tools of Signal Detection Theory*, page 62–96. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Durvasula, K. (2020). Oh gradience, whence do you come? Keynote presentation at the Annual Meeting of Phonology.

Eisner, J. (1997). Efficient generation in primitive optimality theory. In *35th annual meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320.

Ellison, T. M. (1994). *The machine learning of phonological structure*. University of Western Australia.

Ernestus, M. T. C. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language*, 79(1):5–38.

Frisch, S. A., Large, N. R., and Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of memory and language*, 42(4):481–496.

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

Frisch, S. A. and Zawaydeh, B. A. (2001). The psychological reality of ocp-place in arabic. *Language*, pages 91–106.

Fromkin, V. A. (1973). Slips of the tongue. *Scientific American*, 229(6):110–117.

Gallagher, G. (2014). An identity bias in phonotactics: Evidence from cochabamba quechua. *Laboratory Phonology*, 5(3):337–378.

Gallagher, G. (2015). Natural classes in cooccurrence constraints. *Lingua*, 166:80–98.

Gallagher, G. (2016). Asymmetries in the representation of categorical phonotactics. *Language*, pages 557–590.

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.

Goldsmith, J. (1976). *Autosegmental phonology*. PhD thesis, MIT Press London.

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.

Goodman, L. and Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

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.

Guy, G. R. (2007). Lexical exceptions in variable phonology. *University of Pennsylvania Working Papers in Linguistics*, 13(2):9.

Hale, M. and Reiss, C. (2008). *The phonological enterprise*. OUP Oxford.

Haman, E., Etenkowski, B., Łuniewska, M., Szwabe, J., Dąbrowska, E., Szreder, M., and Łaziński, M. (2011). Polish cds corpus. Available from http://childes.psy.cmu.edu.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hayes, B. (2012). Blick: a phonotactic probability calculator (manual).

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., Kobele, G. M., and Riggle, J. (2009). Evaluating the complexity of optimality theory. *Linguistic Inquiry*, 40(2):277–288.

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.

Hua, W., Dai, H., and Jardine, A. (2021). Learning input strictly local functions from their composition. In Reisinger, D. K. E. and Huijsmans, M., editors, *Proceedings of the 37th West Coast Conference on Formal Linguistics*, pages 143–151. Cascadilla Proceedings Project.

Hua, W. and Jardine, A. (2021). Learning input strictly local functions from their composition. In Chandlee, J., Eyraud, R., Heinz, J., Jardine, A., and van Zaanen, M., editors, *Proceedings of the Fifteenth International Conference on Grammatical Inference*, volume 153 of *Proceedings of Machine Learning Research*, pages 47–65. PMLR.

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.

Idsardi, W. J. (2006). A simple proof that optimality theory is computationally intractable. *Linguistic Inquiry*, 37(2):271–275.

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).

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

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.

Jardine, A. and McMullin, K. (2017). Efficient learning of tier-based strictly k-local languages. In *International conference on language and automata theory and applications*, pages 64–76. Springer.

Jarosz, G. (2017). Defying the stimulus: acquisition of complex onsets in polish. *Phonology*, 34(2):269–298.

Jarosz, G., Calamaro, S., and Zentz, J. (2017). Input frequency and the acquisition of syllable structure in polish. *Language acquisition*, 24(4):361–399.

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.

Jusczyk, P. W. and Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., and Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of memory and language*, 32(3):402–420.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of memory and Language*, 33(5):630–645.

Kabak, B. (2011). Turkish vowel harmony. *The Blackwell companion to phonology*, pages 1–24.

Kahng, J. and Durvasula, K. (2023). Can you judge what you don't hear? perception as a source of gradient wordlikeness judgements. *Glossa: a journal of general linguistics*, 8(1).

Kang, Y. (2011). *Loanword phonology*, pages 1–25. Wiley-Blackwell.

Kawahara, S. and Breiss, C. (2021). Exploring the nature of cumulativity in sound symbolism: Experimental studies of pokémonastics with english speakers. *Laboratory Phonology*, 12(1).

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

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.

Lambert, D. and Rogers, J. (2020). Tier-based strictly local stringsets: Perspectives from model and automata theory. *Proceedings of the Society for Computation in Linguistics*, 3(1):330–337.

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.

Lewis, G. L. (2001). *Turkish Grammar*. Oxford University Press, 2nd edition.

Lignos, C. and Gorman, K. (2012). Revisiting frequency and storage in morphological processing. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 48, pages 447–461. Citeseer.

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

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46(1):53–85.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman & Company.

Martin, A. (2011). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language*, 87(4):751–770.

Mayer, C. (2021). Capturing gradience in long-distance phonology using probabilistic tier-based strictly local grammars. *Proceedings of the Society for Computation in Linguistics*, 4(1):39–50.

Mayer, C., McCollum, A., and Eziz, G. (2022). Issues in uyghur phonology. *Language and Linguistics Compass*, 16(12):e12478.

McMullin, K. and Hansson, G. Ó. (2019). Inductive learning of locality relations in segmental phonology. *Laboratory Phonology*, 10(1).

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

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

Moreton, E., Pater, J., and Pertsova, K. (2017). Phonological concept learning. *Cognitive science*, 41(1):4–69.

Nevins, A. and Vaux, B. (2003). Metalinguistic, shmetalinguistic: The phonology of shmreduplication. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 39, pages 702–721. Chicago Linguistic Society.

O'Hara, C. (2020). Frequency matching behavior in on-line maxent learners. *Proceedings of the Society for Computation in Linguistics*, 3(1):463–465.

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

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.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

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

Pierrehumbert, J. (1994). Syllable structure and word structure: a study of triconsonantal clusters in english. *Phonological structure and phonetic form: Papers in Laboratory Phonology III*, pages 168–188.

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

Pierrehumbert, J. (2001b). Why phonological constraints are so coarse-grained. *Language and cognitive processes*, 16(5-6):691–698.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

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

Prince, A. and Tesar, B. (2004). Learning phonotactic distributions. In *Constraints in phonological acquisition*, pages 245–291. Cambridge University Press Cambridge.

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.

Reiss, C. (2017). Substance free phonology. In *The Routledge handbook of phonological theory*, pages 425–452. Routledge.

Rubach, J. and Booij, G. (1990). Syllable structure assignment in polish. *Phonology*, 7(1):121–158.

Scholes, R. J. (1966). *Phonotactic grammaticality*. Mouton & Co.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Shattuck-Hufnagel, S. (1986). The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. *Phonology*, 3:117–149.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

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

Underhill, R. (1976). *Turkish grammar*. MIT press Cambridge, MA.

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

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.

Wolf, M. (2011). Exceptionality. *The Blackwell companion to phonology*, pages 1–23.

Wu, K. and Heinz, J. (2023). String extension learning despite noisy intrusions. In *International Conference on Grammatical Inference*, pages 80–95. PMLR.

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., Lin, I., Yang, M., and Peperkamp, S. (2021). Competition between whole-word and decomposed representations of english prefixed words. *Morphology*, 31:201–237.

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

Zydorowicz, P. and Orzechowska, P. (2017). The study of polish phonotactics: Measures of phonotactic preferability. *Studies in Polish Linguistics*, 12(2).

A Zimmer (1969)'s experiment

In a binary wordlikeness task, Zimmer (1969) asked native adult speakers to select which of two nonce words, for example, temez-temaz, was "more like Turkish". Experiment 1 had 23 participants, and Experiment 2 had 32, all of whom were native adult speakers of Turkish. Table 22 and 23 illustrate the effects of backness and roundness harmony on the wordlikeness experiment carried out by Zimmer (1969). The numbers represent how many participants selected the corresponding nonce word, while the responses indicating "no preference" were excluded.

Experiment 1				Experiment 2			
Harmonic		Disharmonic		Harmonic		Disharmonic	
temez	19	temɑz	3	pemez	30	pemɑz	2
teriz	23	teruz	0	teriz	28	teruz	3
tokɑz	21	tokez	1	tokɑz	26	tokez	6
tipez	21	tipɑz	1	tipez	24	tipɑz	8
teryz	20	teruz	1	teryz	19	teruz	13

Table 22: Effects of backness harmony on Zimmer (1969)'s wordlikeness experiment, from Gorman (2013)

Experiment 1				Experiment 2			
Harmonic		Disharmonic		Harmonic		Disharmonic	
tøryz	19	tøriz	1	pøryz	32	pøriz	0
typyz	22	typiz	0	typyz	31	typiz	1
tɑkuz	15	tɑkuz	3	tɑkuz	22	tɑkuz	10
tɑtiz	12	tɑtuz	6	tɑtiz	20	tɑtuz	12

Table 23: Effects of roundness harmony on Zimmer (1969)'s wordlikeness experiment, from Gorman (2013)

B Polish Training Data

p	4,335	k ʂ	121	t f	42	s p w	19	ʂ w	8	s k f	4	z d ʐ	2	z d l	1
v	2,653	z d	119	m w	42	s t w	18	r v	8	b ʐ d	4	ʑ l	2	f s t ʂ	1
z	2,162	g ʐ	118	x l	40	d͡ʑ	18	m n	8	d b	4	m d͡ʑ d	2	f ʂ t͡ʂ	1
k	2,047	z m	117	d w	40	ʂ l	18	d ɲ	8	k ʂ t	4	z v w	2	m k ɲ	1
m	1,811	ʂ t͡ʂ	116	v ʐ	39	f j	18	c m	8	ʂ r	3	b ʐ m j	2	f s r	1
p ʂ	1,522	ki	115	k f j	38	t͡ʂ t	18	c f	8	p x w	3	c l	2	b d	1
r	1,483	z n	107	s f	38	z v r	17	z g w	8	ʂ p r	3	d v j	2	t k ɲ	1
n	1,389	s m	106	f l	38	z ʐ	17	v z	8	v z g	3	s f j	2	t k f	1
b	1,231	ɕ f j	103	s k w	37	x m	17	d l	8	t r j	3	l ɲ	2	m k n	1
d	1,003	k w	102	g ɲ	37	f t͡ʂ	17	d͡ʑ b	8	r j	3	f c	2	v v j	1
l	911	t ʂ	99	ʂ k	36	z g ɲ	16	t k	7	ʐ v	3	z m ɲ	2	w g	1
t	868	ɕ t͡ʂ	94	z b j	36	d m	16	t j	7	t͡ʂ t͡ʂ	3	s x f	2	v z d w	1
j	773	ɕ l	89	f s	35	ʂ f	15	g n	7	s p s	3	v z b r	2	v z d r	1
g	614	z b	88	ɕ f	34	ʑ r	15	s t f j	7	f t r	3	x f j	2	k m j	1
x	602	z g	86	v l	33	v b	15	x c	7	ʐ ɲ	3	v d͡ʑ	2	s x ɲ	1
s	590	v r	86	x f	33	t͡ʂ f	15	m ʂ	7	d ʐ v	3	s k n	2	g ʑ	1
ɲ	534	z r	82	x t͡ʂ	31	m ɲ	15	b z d	6	ʂ ki j	3	b ʑ	2	z n l	1
ʐ	444	ɕ m j	79	g v	29	s t f	14	b z	6	x j	3	d n	2	b r v	1
s t	427	z j	78	g l	29	z d͡ʑ	14	d͡ʑ v j	6	r d͡ʑ	3	v m j	2	k r t	1
k r	411	z v	78	z l	29	s p j	14	z gi j	6	f s p j	3	k t͡ʂ	2	v gi j	1
w	379	s p ʂ	76	z d r	29	ɕ p	13	d͡ʑ	6	l ʐ	3	v b j	1	v x d͡ʑ	1
d͡ʑ	377	ki j	75	f s k	29	p x	13	x ʂ t͡ʂ	6	ʂ k w	3	b z v	1	z m r	1
f	375	s p r	74	m l	29	f ɕ t͡ʂ	13	s ʂ	6	v z g l	3	l j	1	t͡ʂ m j	1
t͡ʂ	370	k ɕ	74	d͡ʑ v	28	f k r	12	v z r	6	t͡ʂ k	3	t ʂ n	1	x ʂ t͡ʂ	1
s p	370	b l	73	f s p	28	f p r	12	k n	6	t͡ʂ m	3	v z ɲ	1	v d r	1
p j	365	z v j	68	s k l	28	z gi	12	t͡ʂ f j	6	f s t r	3	w b	1	p s t	1
v j	344	f s t	67	k t	28	gi j	12	z g ʐ	6	ʂ t r	3	z v l	1	ʂ p t͡ʂ	1
p r	340	z m j	67	f ɕ	27	s c	12	g z	5	v b r	2	f p j	1	k m	1
ʂ	338	f r	67	f ʂ	27	f t	12	f k l	5	t n	2	g ʐ j	1	x m j	1
c	331	s k ʂ	62	p ɕ	27	p ʂ t͡ʂ	12	f p w	5	l n	2	ʂ j	1	v z n	1
t͡ʂ	327	s x	60	p t	27	t͡ʂ w	12	w z	5	g m	2	z z	1	v z b	1
ɕ	305	f p	59	v ʑ	25	t x	11	k l j	5	p x n	2	ɕ w	1	w b j	1
t r	266	d v	58	f t͡ʂ	25	ɕ m	11	t ʂ t͡ʂ	5	b r v j	2	f x ʂ	1	v g	1
s k	257	ɕ p j	57	ʂ n	24	x ʂ	11	z d m	5	m g l	2	t͡ʂ m	1	t͡ʂ l	1
g r	249	p s	54	g v j	24	v d	11	f s x	5	s m r	2	ʂ p j	1	t͡ʂ j	1
m j	248	b w	52	ʂ k l	24	g d͡ʑ	10	c n	5	g ʐ m	2	m ʐ	1	v v	1
s w	227	s t ʂ	51	ʂ t	23	z b r	10	f k	4	d ʐ v j	2	ʐ j	1	s x w	1
b r	206	b ʐ	51	ʂ m	23	p s t r	10	p ɲ	4	v m	2	ʐ l	1	x ʂ t	1
k l	196	x w	49	d ʐ	23	s s	10	s t͡ʂ	4	r t	2	ʐ n	1	n z	1
d r	190	ʂ p	48	t͡ʂ f	22	s ɕ	10	s t͡ʂ	4	g ʐ b j	2	ʐ b	1	d͡ʑ d͡ʑ	1
p w	175	z ɲ	47	m r	22	t l	9	m g w	4	t͡ʂ x	2	ʐ r	1	v g ɲ	1
p l	172	t w	46	z d j	20	t r v	9	ʐ m	4	r ʐ	2	g d͡ʑ	1	b ʐ m	1
v w	146	gi	46	s r	20	ʐ w	9	z b l	4	s p l	2	s ɲ	1	p t͡ʂ	1
z w	140	ɕ ɲ	46	v n	20	d͡ʑ v	9	t ɲ	4	d r v	2	t ʂ m j	1	f ɕ r	1
ʑ	139	x r	46	d j	20	s l	9	t f j	4	k r v j	2	z b w	1	d z	1
s t r	137	k f	46	f x	20	s n	9	k r v	4	m x	2	m d l	1	t s t	1
b j	133	ɕ r	44	g d	19	z g r	9	v z m	4	l ɕ ɲ	2	v ɲ	1	v g w	1
g w	121	s k r	44	f k w	19	l v	8	s x r	4	r ʐ n	2	l v j	1	r ʐ ɲ	1