Phonology (to appear), **xx**:xx 1–46
doi:xxxxxxxxxxxxxxxx

PHONOLOGY

ARTICLE

An Exception-Filtering Approach to Phonotactic Learning

Huteng Dai[1,2]

[1]Department of Linguistics, Rutgers University, New Brunswick, USA.
[2]Department of Linguistics, University of Michigan, Ann Arbor, USA Email: huteng@umich.edu .

Received: 08 February 2023; **Accepted:** 26 December 2023

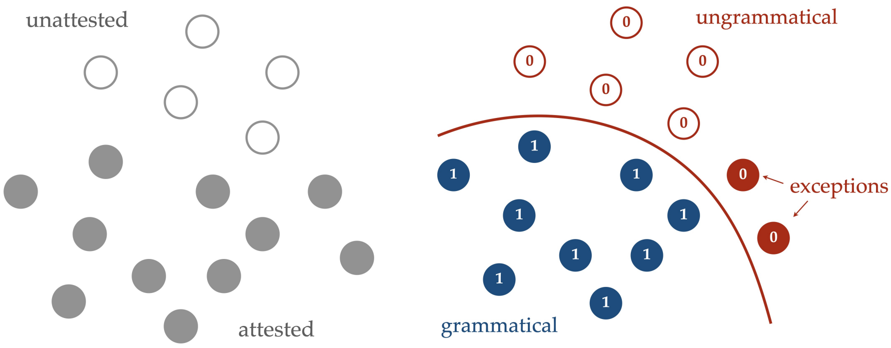
Keywords: phonotactics, phonological learning, categorical grammar, exceptionality, indirect negative evidence, frequency, onsets, vowel harmony, acceptability, English, Polish, Turkish

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" approach that harnesses the discrete nature of categorical grammars to filter out lexical exceptions using statistical criteria adapted from probabilistic models. Applied to naturalistic corpora from English, Polish, and Turkish, the learnt grammars showed a high correlation with the acceptability judgements in behavioural experiments. Compared to benchmark models, the model performed increasingly better with data that contain a higher proportion of lexical exceptions, reaching its peak in learning Turkish nonlocal vowel phonotactics, highlighting its ability to handle lexical exceptions.

1. Introduction

There exist logically infinite potential sound sequences in any given language, yet only some are considered well-formed by speakers. *Phonotactics* refers to this implicit knowledge of speakers to discern well-formed sound sequences in their language, which does not apply uniformly to the entire lexicon—certain lexical exceptions can violate otherwise universally applicable patterns (Guy, 2007; Wolf, 2011). However, children can acquire regular patterns in the presence of lexical exceptions. For example, despite the existence of disharmonic sequences in their language experience, experimental studies have shown that Turkish infants tune into nonlocal phonotactics in vowel harmony patterns as early as six months (Altan *et al.*, 2016; Hohenberger *et al.*, 2016; Sundara *et al.*, 2022, see §7 for details).

The challenge of phonotactic learning in the presence of lexical exceptions is illustrated in Figure 1. Under the positive evidence-only assumption, the learner relies exclusively on unlabelled input data (Marcus, 1993), denoted by the filled dots in the figures; conversely, the unfilled 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 sequences, represented by 1s and 0s in Figure 1b.



(a) Potential sound sequences.

(b) Target grammar.

Figure 1: The learning problem in the presence of exceptions (adapted from Mohri *et al.* (2018: 8)). In both (a) and (b), filled dots represent attested data, while unfilled dots indicate unattested data. In (b), 0 indicates the ungrammatical items and 1 indicates grammatical items, assuming a boolean grammaticality.

Learning models that assume all attested sound sequences as grammatical data run the risk of building attested but ungrammatical 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). 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 & 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 and ungrammatical sequences (Yang 2016: 3), which can facilitate the identification of lexical exceptions. However, learning models based on categorical grammars are generally considered vulnerable to exceptions in naturalistic corpora, as discussed in Gouskova & Gallagher (2020, added emphasis and adapted spelling):

[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.

> "In contrast to our approach, Heinz (2010), Jardine (2016), and Jardine & 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 & Gallagher, 2018)."

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 & 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 a 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 Hayes & Wilson (2008) phonotactic learner and traditional O/E criterion (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 discrete nature of categorical grammars allows the proposed model to completely filter out lexical exceptions and demonstrates robust performance across noisy corpora from English, Polish, and Turkish, successfully learning phonotactic grammars that approximate acceptability judgements in behavioural experiments. Compared to benchmark models, the model performed increasingly better with data that contain a higher proportion of lexical exceptions, reaching its peak in learning Turkish nonlocal vowel phonotactics, despite the complexity introduced by disharmonic forms in the 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 the 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. The relationship between these components is visualised in Figure 2. Together, the lexicon and grammar form the *competence*, representing the internalised knowledge. Speakers' acceptability judgments are influenced by both competence and performance factors. For example, a word [sfid] will receive low acceptability if the lexicon does not consist of the word and the grammar penalizes its substructure *sf. As highlighted in Figure 2, this paper focuses on the acquisition of grammar, abstracting away from lexicon acquisition and general performance factors.

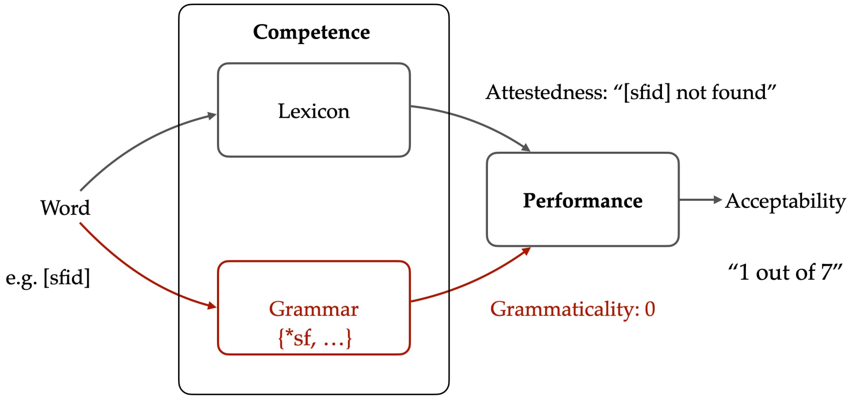


Figure 2: The relationship between lexicon, grammar, and performance. The rating "1 out of 7" is provided as an example and does not represent actual data.

The current study distinguishes between the terms *grammaticality* (or well-formedness) and *acceptability*, which have frequently been conflated in previous research in phonology (Hayes & Wilson, 2008; Albright, 2009). In this context, acceptability refers to the judgements made by 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. Similar to the *dual-route* model (Pinker & Prince, 1988; Zuraw, 2000; Zuraw *et al.*, 2021), the lexical route allows the speaker to access the lexicon and evaluate the acceptability of existing (or *attested*) words, regardless of possible grammar violations. If the lexicon does not contain certain sound sequences, as in nonce words, the speaker instead evaluates their acceptability in the grammar via the non-lexical route, in which grammaticality is predicted based on grammar. This grammaticality then interacts with other extragrammatical factors and results in the acceptability in the performance level.

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. Due to the existence of extragrammatical factors, models that perfectly align with acceptability could actually deviate 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 the grammar (Kahng & Durvasula 2023: 3).

Acceptability judgements are commonly collected via rating tasks employing a numeric Likert scale and characterised as "gradient" (non-categorical) in nature (Albright, 2009). Individual Likert ratings correspond to categorical multilevel, 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.[2] 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 judgement is traced back to *auditory illusions*, as shown in Kahng & Durvasula (2023).

In light of these considerations, the acceptability judgements reported in previous studies are not incompatible with categorical grammar. On the one hand, the current study assumes that the grammaticality of sound sequences, categorical or probabilistic, is *reflected* in acceptability judgements, and a successful grammar should exhibit a robust correlation between predicted grammaticality and acceptability judgements to allow "direct investigation" of linguistic competence (Lau *et al.*, 2017). On the other hand, the current study argues that gradient acceptability collected through numerical rating tasks does not necessitate gradient / probabilistic grammars nor negate categorical grammars (cf. Coleman & Pierrehumbert, 1997; Hayes & Wilson, 2008).

The current study employs categorical grammars using a discrete set of constraints that simply accept grammatical sequences and reject ungrammatical ones. On the contrary, probabilistic grammars, such as Maximum Entropy (MaxEnt) grammars (Hayes & 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. However, categorical grammars cannot be conflated with probabilistic grammars with thresholds (Hale & Smolensky, 2006), which cannot define infinite languages, as mathematically demonstrated in Alves (2023). Probabilistic grammars have been noted for their ability to model human sensitivity to frequency information and approximate gradient acceptability judgements (Hayes & Wilson, 2008), while categorical grammars delineate a clear boundary between grammatical words and lexical exceptions (Yang 2016: 3). This discrete nature can be used to facilitate phonological learning, as shown in the current study.

Hale & Reiss (2008: 18) adopted a nihilistic view of phonotactic grammars, arguing that phonotactics is not part of phonological grammar, as it is computationally inert

[2] 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 judgements.

in morphophonological alternations (Reiss 2017: §6). However, experimental evidence has shown that infants acquire morphologically agnostic phonotactics, and the learnt phonotactics can facilitate the learning of morphophonological alternations (Jusczyk *et al.*, 1993, 1994; Jusczyk & Aslin, 1995; Archer & Curtin, 2016; Chong, 2021). Gorman (2013: §1) also demonstrated internalisation of phonotactic constraints in various domains, such as wordlikeness judgements and loanword adaptation. Furthermore, the current study upholds the concept of categorical grammars, which essentially motivated the adoption of the nihilistic view (Reiss 2017: 14; "categorical baby"). In light of this, the current study models the learning of phonotactic grammar as a crucial component within a broader framework of phonological learning (see the discussion in §8).

2.2. Attestedness vs Grammaticality

While the "grammar" acts as a finite system representing an infinite number of grammatical sound sequences, the term "lexicon" refers to all words that speakers know, including all exceptional and unpredictable features of attested input data (Chomsky 1965: 229; Chomsky & Halle 1965; Jackendoff 2002: 153). In turn, the input data in phonotactic learning drawn from the lexicon can include sound sequences that deviate from grammar.

The current study assumes that the exceptionality is not a *static* label in the input data but emerges from the discrepancy between *attestedness w.r.t.* the input data and *grammaticality w.r.t.* the hypothesis grammar. Grammaticality indicates whether phonological representations conform to the hypothesis grammar internalised by the learner. Researchers have used various converging methodologies to approximate the hypothesis grammar, especially statistical generalisations (e.g., observed-to-expected ratio; detailed in §3) or performance data such as nonce word acceptability (detailed below) and speech errors.[3] For convenience in the discussion, consider a hypothesis grammar consists of categorical constraints {*sf, *bn}. The symbol * is only used to indicate ungrammatical sequences (as opposed to unattested). In contrast, attestedness indicates whether a sound sequence occurs in the input data. [brɪk] (as in *brick*) and [*sfiɚ] (*sphere*) are both attested in the English lexicon, while [blɪk] (*blick*) and [*bnɪk] (*bnick*) are not, as illustrated in Table 1.

	grammatical	**ungrammatical**
attested	[**br**ɪk]	[***sf**iɚ]
unattested	[**bl**ɪk]	[***bn**ɪk]

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

[3] Speech errors elicitation have been used to probe phonotactic constraints (Fromkin, 1973), such as nonlocal consonant cooccurrences (Rose & King, 2007).

This discrepancy between attestedness and grammaticality yields both accidental gaps (grammatical but unattested) and lexical 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 speakers uniformly reject **bnick* while accepting *blick*, a classic example of accidental gaps (Chomsky & Halle, 1965; Hayes & Wilson, 2008).

The attested sequences are considered ungrammatical lexical exceptions if and only if they violate the hypothesis grammar, such as {*sf, *bn} in the above example. *Sphere* is a classic example of lexical exceptions: the onset [sf] rarely occurs in English and has been labelled ungrammatical in previous work (Hyman, 1975; Algeo, 1978; Kostyszyn & Heinz, 2022). The architecture in Figure 2 predicts that the acceptability of the attested word "sphere" itself is directly influenced by the lexicon and is considered highly acceptable by some speakers. However, when they are not stored in the lexicon, [sf]-onset nonce words are commonly judged unacceptable, as shown in an experiment conducted by Scholes (1966: 114): 33 seventh-grade English speakers were asked if a nonce word "is likely to be usable as a word of English." Only 7 participants responded "yes" to the [sf]-onset nonce word [**sf**id], lower than [**bl**ʊŋ] (31 "yes"), and even lower than words with unattested onsets such as [**ml**ʊŋ] (13 "yes"). Leveraging the converging evidence that *sf is a phonotactic constraint in hypothesis grammar, the attested [sf]-onset word *sphere* can be considered as a lexical exception, in contrast to attested and grammatical *brick*.[4]

Lexical exceptions are also 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*. 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) and receive relatively low acceptability scores in nonce word judgements (Scholes, 1966, also see §5). 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 input data drawn from the lexicon can contain ungrammatical lexical exceptions according to the hypothesised phonotactic grammar.

2.3. Summary

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

[4] [br]-onset nonce words are not included in Scholes's experiment, but they are rated more acceptable than in [bl]-onset nonce words in Daland *et al.* (2011), as shown in §5.

3. Exception-Filtering Phonotactic Learner

This section proposes a “categorical grammar + exception-filtering” approach to select a hypothesised categorical grammar (hereafter “hypothesis grammar”) from the hypothesis space. This 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 primary objective 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 the computational level (Marr, 1982). For this reason, the current proposal adopts segmental representations derived from input data for its practical advantages, a departure from prespecified feature representations advocated by previous studies (Hayes & Wilson, 2008; Gouskova & Gallagher, 2020). 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, when training data are phonemically transcribed, 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

From a constraint-based view of grammar, phonotactic learning involves selecting a hypothesis grammar (G; a set of constraints) from the hypothesis space (Con; adapted from the OT terminology). 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 & Pierrehumbert, 1997; Breiss, 2020; Kawahara & 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 & Smolensky, 1993; Prince & Tesar, 2004). In contrast to OT, the hypothesis grammar in the current proposal is drawn from a highly restrictive hypothesis space.[5] 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 & Heinz, 2016; Lambert & 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 strings (e.g.,

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

sound sequences) that adhere to its associated grammar, which can be mathematically 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 & 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 local constraint *øi on the vowel tier. This concept, although similar, is distinct from the traditional feature-based definition in Autosegmental Phonology (Goldsmith, 1976)

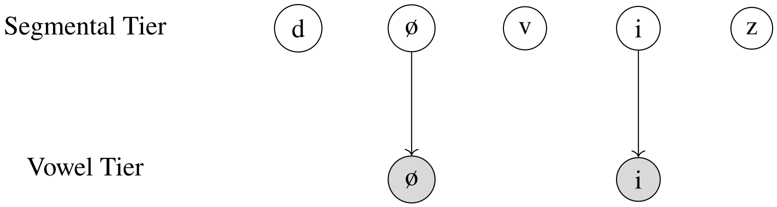


Figure 3: Extraction of vowel tier from the Turkish word [døviz] "currency". The vowel tier contains the vowels in this word, disregarding the non-tier consonants.

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, $\texttt{factor}\,(\mathrm{CCV}, 2) = \{\mathrm{CC}, \mathrm{CV}\}$, and $\texttt{factor}\,(\mathrm{CVC}, 2) = \{\mathrm{CV}, \mathrm{VC}\}$. 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 $G = \{*\mathrm{CC}\}$, 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 but typologically robust hypothesis space, capturing a range of local and nonlocal phonotactics (Heinz *et al.*, 2011). Specifically, McMullin & Hansson (2019) provides experimental evidence for TSL_2 as a viable working hypothesis space for phonotactic learning, demonstrating that adult 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 & Heinz, 2016; Jardine &

McMullin, 2017). This approach has been successfully applied in previous work spanning both probabilistic and categorical approaches (Hayes & Wilson, 2008; Gouskova & Gallagher, 2020; Mayer, 2021; Dai *et al.*, 2023; Heinz, 2007; Jardine & Heinz, 2016).

One of the main challenges of phonotactic learning, as mentioned in Hayes & 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 & Gallagher, 2020), Hungarian vowel harmony (Hayes & Londe, 2006), and Arabic OCP-Place patterning (Frisch & 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 and O/E Criterion*

The goal of phonotactic learning 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 grammar. 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. An in-depth review of this issue can be found in Wu & Heinz (2023).

One approach to address the challenge of exceptions utilises an *exception-filtering* mechanism to exclude exceptions while learning categorical grammars. Hayes & Wilson (2008: 427-428) hypothesised that children possess an innate ability to discern the unique status of certain exotic items and improved their learning results by excluding exotic items from input data. This ability to detect and exclude anomalies aligns closely with the concept of exception-filtering in the current proposal. Although such a mechanism was considered challenging to propose (Clark & Lappin 2010: 105), the current study achieves this by leveraging *indirect negative evidence* derived from frequency information (Clark & Lappin, 2009; Pearl & Lidz, 2009; Yang, 2016), specifically from type frequency (Pierrehumbert, 2001; Hayes & Wilson, 2008; Richtsmeier, 2011).[6] Indirect negative evidence allows learners to infer grammaticality labels from unseen data, despite the absence of such labels in positive evidence, guided by the principle

[6] 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 & Wilson 2008: 395). The current study leaves this alternative strategy for future investigation.

that a sequence that occurs less frequently than expected in the input data is likely ungrammatical.

The comparison between observed (O) and expected (E) type frequencies embodies the exception-filtering mechanism in the current study and has been widely applied in the identification of phonotactic constraints (Pierrehumbert, 1993, 2001; Frisch *et al.*, 2004; Hayes & Wilson, 2008) since Trubetzkoy (1939: Chapter VII). 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.

The traditional O/E equation proposed by Pierrehumbert (1993) has been widely applied to discover phonotactic constraints (Pierrehumbert, 2001; Frisch *et al.*, 2004). However, this equation assumes an empty hypothesis grammar, which becomes inaccurate once any constraint is added, as discussed in Wilson & Obdeyn (2009) and Wilson (2022).

The current criterion O/E draws inspiration from Hayes & Wilson (2008), while a crucial difference lies in the definition that the hypothesis grammar in the current study is noncumulative, leading to distinct calculations of 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:

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

In a toy sample $S = \{\text{CVC, CVV, VVC, VVV, VCV, CCV}\}$, $O\left[{}^*\text{CC}\right] = 1$, $O\left[{}^*\text{CV}\right] = 4$, $O\left[{}^*\text{VC}\right] = 3$, $O\left[{}^*\text{VV}\right] = 3$. Here, $O\left[{}^*\text{VV}\right]$ is 3 rather than 4 because, by definition, $O\left[C\right]$ counts the number of strings violated by the potential constraint (at least once) rather than the cumulative number of substring violations across all strings. Therefore, $O\left[{}^*\text{VV}\right]$ only counts once in the string VVV. 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 $E\left[C\right]$ represents the number of unique strings in the hypothesised language L that violate C, under a noncumulative hypothesis grammar G.[7] Following Hayes & Wilson (2008), the current study works with an estimation to $E\left[C\right]$ by limiting the maximum string length in L to $\ell_{\max}$, mirroring the length of the longest string in the input data S. $E\left[C\right]$ is then approximated by:

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

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

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

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

$\mathit{Ratio}\,(C, G, \ell)$ 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]

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

$\mathit{Count}\,(G, \ell)$ is the count of unique ℓ-length strings in the hypothesis language L accepted by G. Therefore, $\mathit{Count}\,(G, \ell) - \mathit{Count}\,(G', \ell)$ 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 $|S_3| = 8$). $\mathit{Count}\,(\emptyset, 3) = 8$, 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.

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

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

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.

[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.

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 ∈ G). In other potential constraints, for example, E [*VV] $= |S_3| \cdot \left(\frac{5-2}{5}\right) = 5 \cdot \frac{3}{5} = 3$, 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 ($O = E$). During the learning process, a constraint is included in the grammar if the O/E ratio falls below a specified threshold ($O/E < \theta$). This comparison is performed at increasing threshold levels, ranging from 0.001 to $\theta_{\max}$, also known as the *accuracy schedule* (Hayes & Wilson, 2008), where the interval after 0.1 is fixed to 0.1. For example, the accuracy schedule $\Theta = [0.001, 0.01, 0.1, 0.2, 0.3, \ldots, 1]$ if $\theta_{\max} = 1$. This structure prioritises the integration of potential constraints with the lowest O/E values.[9] $\theta_{\max}$ can be interpreted as follows: the higher $\theta_{\max}$ indicates the need for more statistical support, i.e. higher O/E, before considering a two-factor as grammatical. This also allows for the modelling of individual variability in phonotactic learning, where some learners require more statistical support for grammatical sequences, reflected by a higher $\theta_{\max}$.

Equipping the Exception-Filtering learner with the accuracy schedule adapted from Hayes & Wilson (2008) controls the contrast between them and facilitates direct comparison between their best-performing models. Dealing with realistic corpora and experimental data requires posterior adjustments of $\theta_{\max}$: the analyst/user sets this hyperparameter to the value between 0 and 1 that achieves the highest scores on all statistical tests in each test dataset. In this paper, $\theta_{\max}$ is set to 0.1 for the English and Polish case studies and 0.5 for the Turkish case study. The current study shows that once an appropriate hyperparameter is in place, the proposed model can successfully acquire categorical grammars despite the existence of lexical exceptions.

Future psycholinguistic studies are required to better model the factors that determine $\theta_{\max}$. For example, Frisch *et al.* (2001) showed that the larger the *lexicon size* of individual participants in their experiment, the more likely they would accept sequences with low type frequency, which means lower $\theta_{\max}$ in the Exception-Filtering learner.

3.4. *Learning Procedure*

Building on the concepts above, the Exception-Filtering learner models how a child learner acquires 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

[9] The current proposal leverages the normal approximation technique to refine the O/E ratio, applying a statistical upper confidence limit (UCL) given by $p + \sqrt{\frac{p(1-p)}{n}} \times t^{(n-1)}_{(1-\alpha)/2}$, where p is the O/E ratio, n is the sample size (proportional to E value), and $t^{(n-1)}_{(1-\alpha)/2}$ the t-value for a two-tailed test at significance level α with $n - 1$ degrees of freedom (Mikheev, 1997; Albright & Hayes, 2002, 2003; Hayes & Wilson, 2008). α is set to 0.975 after Hayes & Wilson (2008). This adaptation provides more nuanced differentiation in O/E evaluations, especially prominent between figures such as 0/10 and 0/1,000, resulting in UCLs of 0.22 and 0.002, respectively. This differentiation helps to prioritise potential constraints where the O and E disparity is high.

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 reserved for future investigations.

Let us look at a toy example: given the tier (also the inventory) {C, V}, consider the target grammar $\mathcal{T} = \{*\mathrm{CC}\}$. 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, the learner can handle samples with varying lengths.

As visualised in Figure 4, 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 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 termination. The following paragraphs illustrate this learning procedure using the toy input data with the exception of *CCV. Given the page limitations, a simplified accuracy schedule $\Theta = [0.5, 1]$ with $\theta_{\text{max}} = 1$ is used to avoid too many iterations.

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.

[10] The assumption that a single uniform target grammar applies to all 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 speakers.

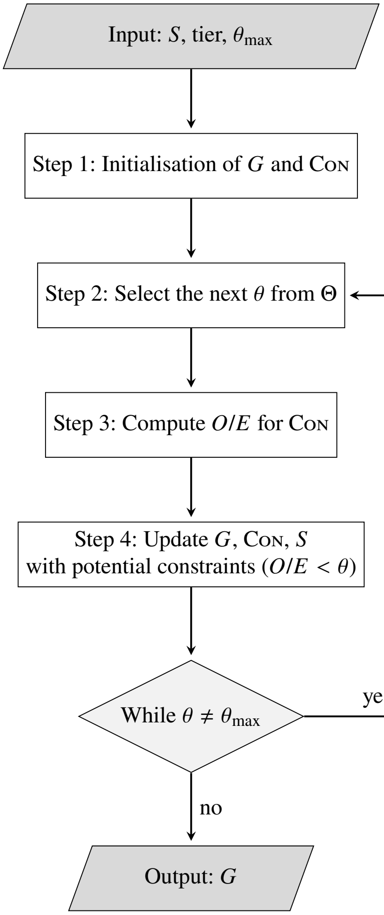

Input: S, tier, θ_{max}
Step 1: Initialisation of G and Con
Step 2: Select the next θ from Θ
Step 3: Compute O/E for Con
Step 4: Update G, Con, S with potential constraints ($O/E < \theta$)
While $\theta \neq \theta_{\text{max}}$
yes
no
Output: G

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

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 in the language defined by G (namely, CCV, VCC, CCC) violate the updated grammar $G' = \{*\text{CC}\}$. The ratio that a string violates *CC in the sample is

$\mathit{Ratio}\,(\text{*CC}, \emptyset, 3) = 1 - \frac{5}{8} = \frac{3}{8}$. As a result, $E\,[\text{*CC}] = |S| \cdot \mathit{Ratio}\,(\text{*CC}, \emptyset, 3) = 6 \cdot \frac{3}{8} =$ 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.[11] 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 learning process without the negative impact of identified lexical exceptions.

3.4.4. Iteration and Termination

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

[11] This filtering mechanism does not exist in Hayes & 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.

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 termination of learning. The term *convergence* is avoided in this context because establishing its conditions requires a more general proof, which is reserved for future research.

In the second iteration of the toy example, after *CC is added to G and removed from Con (hence "-" in the O [*CC] and E [*CC] of Table 6), θ is reassigned to 1, and no constraint satisfies $O/E < \theta$. $\theta = \theta_{\text{max}} = 1$ indicates the termination of the learning process. The learnt grammar matches the target grammar $\mathcal{T} = \{\text{*CC}\}$, as shown in Table 6.

.

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

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

Table 6: Step 2 and 3 after the first iteration.

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 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.

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. Statistical assessment: evaluating the performance of the model with new data. This is achieved through the statistical tests against test dataset as discussed below;
4. Statistical 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 and nonlocal vowel phonotactics in Turkish. Learning from onset phonotactics controls the influence of syllable structures and considerably simplifies the learning problem (Daland *et al.*, 2011; Jarosz, 2017; Jarosz & 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 & 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; similar to θ_{max} in this paper) and Max gram size n (2 to 3) in the HW learner were fine-tuned so that only the highest performing models across all tests are reported.[12] 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 & Wilson 2008: 387; $\mu = 0, \sigma = 1$; see more discussion in §8).[13]

The current study also implements a simple categorical Tier-based Strictly 2-Local phonotactic learner (henceforth Baseline; capitalised to distinguish from other baseline models), adapted from *memory-seg* learner (Wilson & Gallagher, 2018) and other previous work (Gorman, 2013; Kostyszyn & 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 judgements 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 judgements using the experimental data collected by Zimmer (1969). This is in line with the methodology employed by Hayes & Wilson (2008) for deriving acceptability judgements 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 example, Hayes & Wilson (2008) compared the learnt grammars of Shona and Wargamay with the phonological generalisations in the previous literature. Gouskova & Gallagher

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

[13] The current study omits the complementation operator $^\wedge$, which introduces implicational constraints in Hayes & Wilson (2008: §4.1.1). For example, $\left[{}^\wedge \alpha F, \beta G, \ldots\right]$ denotes any segment that is not a member of the natural class $[\alpha F, \beta G, \ldots]$. The constraint that limits pre-nasal segments to [s] would be formulated as *[$^\wedge$-voice, +ant, +strident][+nasal] "if the segment precedes [+nasal] is not [s], assign a violation". For the current case studies, this operator 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. Omitting this additional mechanism controls the difference in comparison with other models that do no employ complementation operator.

(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 statistical assessment and comparison are described below:

4.1. Correlation Tests

The correlation between predicted judgements and gradient acceptability judgements, 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 & 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 & 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, Pearson's r is not reported in this study.[14]

Non-parametric tests measuring rank correlations are more appropriate as they make weaker assumptions about the distribution of acceptability judgements (Gorman 2013: 27). Spearman's ρ assumes monotonicity, meaning that the lower values in acceptability consistently correspond to lower levels of predicted grammaticality score. Spearman's ρ requires stronger monotonic relationships to produce higher correlation coefficients, making the score more sensitive to inconsistent performance of subjects, compared to other non-parametric tests. For example, if subjects inconsistently assign ratings from a scale of 1-to-6 to intermediate judgements, where a score of 4 could represent less or equal grammaticality as a score of 2, this will disrupt monotonicity and thus greatly lower Spearman's ρ.

In Goodman-Kruskal's γ and Kendall's τ test, pairs of observations (X_i, Y_i) and (X_j, Y_j) from predicted judgements (X) and gradient acceptability judgements (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: $\gamma =$ (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.

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

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 judgements (Hayes & 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. The current study will evaluate this thresholding method empirically, while Alves (2023) has mathematically and theoretically shown the consequences of probabilistic grammars with thresholds.

The following three sections employ the methodologies described above to 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 & Hayes (2003) and Scholes (1966). This section 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

To facilitate comparison with previous work, this case study uses a "modestly" exceptionful data in Hayes & Wilson (2008, appendix B) the input data, assuming that this dataset has a similar distribution of type frequency as in children's learning experience. The dataset consists of 31,985 onsets taken from distinct 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; Daland *et al.*, 2011; Hayes, 2012). This provides a representative sample that approximates the type frequencies of English onsets in the language experience of English speakers.

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	ʤ	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	ʧ	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.

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 & 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 & 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].[15] This filtering of [Cj] onsets leads to the input data characterised as "modestly exceptionful" because there are only few remaining exotic onsets.

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 judgements. 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).[16] 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 & Rysling, 2017), would require feature-based 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 23 consonants that appear in the input data based on the TSL_2 language, excluding phonemes that never occur at word

[15] Gorman (2013: 98) reviewed several empirical evidence for this analysis. For example, [ju] behaves as a unit in language games (Davis & Hammond, 1995; Nevins & Vaux, 2003) and speech errors (Shattuck-Hufnagel 1986: 130).

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

initial positions such as [x] (as in *loch*) and [ŋ] (*ring*). As a result, the hypothesis space is populated with a total of $23 * 23 = 529$ 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.

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 best performance during the evaluation. The leftmost column displays the first symbol in a two-factor, while the second 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 & 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 & Wilson, in which these two-factors are treated as violable constraints.

5.3. *Model Evaluation in English*

This section evaluates whether the learnt grammar approximates the acceptability judgements 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.

	Stops						Affricates		Fricatives									Nasals		Liquids		Glides	
	p	t	k	b	d	ɡ	ʧ	ʤ	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
ɡ	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
ʧ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ʤ	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 displayed on the leftmost column, while the second symbols are represented by segments on the second top row. The highlighted cells indicate the attested two-factors within the input data, while blue and red separately correspond to grammatical and ungrammatical two-factors.

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).

Table 10 provides a performance comparison among the Exception-Filtering ($\theta_{\max} = 0.1$), Baseline, and HW learner (Max $O/E = 0.3$, Max gram $n = 3$, the same as Hayes & 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].

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

		Exception-Filtering	Baseline	HW
Correlation (Overall)	Spearman's ρ	0.834	0.839	$\underline{0.931}$
	Goodman-Kruskal's γ	0.996	$\underline{1}$	0.860
	Kendall's τ	0.690	0.693	$\underline{0.8}$

Table 10: Results of the best performance in Exception-Filtering ($\theta_{\max} = 0.1$), Baseline, and HW learner (Max $O/E = 0.3$, $n = 3$); 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,[17] 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.

[17] The only difference is that Exception-Filtering learner can learn constraint *ʃl which receives an intermediate 3.125 averaged Likert rating, while the Baseline learner cannot.

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 judgements 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 datasets.

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 judgements from the behavioural data in Daland *et al.* (2011), and managed to deliver a robust performance comparable to benchmark HW model 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 & 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.

	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	k^j, g^j					

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 & Rysling, 2017; Kostyszyn & Heinz, 2022). Polish is known for allowing complex onsets (up to four consonants such as [vzdw]) that defy SSP (Jarosz, 2017; Kostyszyn & Heinz, 2022).[18] 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.

	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 grammar from Polish input data. The first symbol of a two-factor sequence is displayed on the leftmost column, while the second symbols are represented by segments on the second 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.

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

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 best performance during the evaluation. The learnt grammar provides intriguing information on 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 & 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 (having a low or equal sonority rise), namely [lv, rv, mn, mɲ]. In essence, while a comprehensive evaluation 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.

6.3. Model Evaluation in Polish Data

This section evaluates the degree to which the learnt grammar reflects acceptability judgements 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 & Rysling, 2017).

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.[19] For instance, [ʐj] 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 the average Likert ratings in Table 13. [20] 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 (Max O/E = 0.7, n = 2), with a modestly lower Spearman's ρ and a modestly higher Kendall's τ.

The Exception-Filtering learner identified more exceptional two-factors within the Polish input data. Moreover, its performance relative to the benchmark models

[19] 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.

[20] The correlation scores are compared across the entire test dataset as a whole, rather than separately for attested (type frequency $>$ 0) and unattested (type frequency $=$ 0) sequences as in Jarosz & 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 nullifies the correlation tests in unattested sequences.

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

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

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 summarise, the Exception-Filtering learner, trained on Polish child-directed corpus, has illustrated its potential in extracting categorical grammars that approximate acceptability judgements. The performance of the model is on par with the HW learner in Spearman's ρ, and 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 judgements. These results further substantiate the potential of the Exception-Filtering learner in inducing phonotactic patterns from realistic corpora.

		Exception-Filtering	**Baseline**	**HW**
Correlation (Overall)	Spearman's ρ	0.789	0.712	$\underline{0.808}$
	Goodman-Kruskal's γ	$\underline{0.958}$	0.823	0.639
	Kendall's τ	$\underline{0.651}$	0.586	0.640

Table 14: Results of the best performance in Exception-Filtering ($\theta_{\text{max}} = 0.1$), Baseline, and HW learner (Max $O/E = 0.7$, $n = 2$); correlation tests are approximating averaged Likert ratings in Polish; categorised based on attestedness; best scores are underscored.

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 a comprehensive 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/ "stamp", [lɑr] instead of [ler] surfaces "stamps".

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 & 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.[21]

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.

[21] 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 speakers, as shown in the ratings of nonce words by 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).

Generally, a substantial number of exceptions to these phonotactic patterns arise from compounds and loanwords (Lewis, 2001; Göksel & 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 & Kerslake (2004: 11), and Kabak (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 $\approx 66,000$

roots and the elicited derived forms (root + affixes) produced by two adult native Turkish speakers.[22] 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 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 phonotactic 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	3,216	202	1,000
e	15,984	7,130	591	129	663	2,873	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	10,267	1,526	1,656
u	475	606	147	40	153	3,035	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 eight 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.[23] 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* speakers. However, these categorical labels are supported by Zimmer (1969)’s behavioural experiment.

[22] 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 & Gorman, 2012).

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

In a binary wordlikeness task, Zimmer (1969) asked Turkish native adult speakers to select which of two nonce words (e.g., temez-temaz) was “more like Turkish”. Experiment 1 had 23 participants and Experiment 2 had 32 (see appendix A for details); the majority of participants preferred the harmonic to disharmonic roots in a yes/no rating task, which provides 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 judgements of the *majority* of participants in Zimmer (1969)’s experiment. This follows the common practise in previous computational studies when acceptability judgements of nonce words in the test dataset are not accessible. For example, Gouskova & 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	0.969	0.238	0.906
	ungrammatical	1	0	0.875
	grammatical	0.875	1	0.917
	F-score	0.933	0.385	0.824
	precision	1	0.238	0.778
	recall	0.875	1	0.875

Table 19: Performance comparison of Exception-Filtering ($\theta_{\max}$ = 0.5), Baseline, and HW learner (Max O/E = 0.7, n = 3) 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.

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 degrees of shading is proportional to the negative harmony scores, which is rescaled according to the minimum and maximum harmony score.

$\sigma_1 \downarrow \sigma_2 \rightarrow$	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
ɑ	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.

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.[24]

7.3.2. The Second Test Dataset (Approximated Acceptability Judgements)

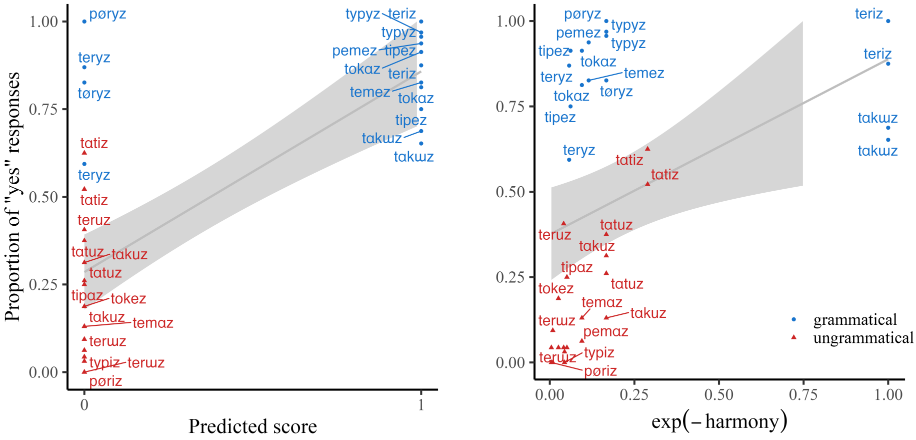
The purpose of the second test dataset is to demonstrate that the learnt categorical grammar can approximate the acceptability judgements in the behavioural data. The second testing data includes 36 nonce words in Zimmer (1969), and takes the proportion of "yes" responses averaged across participants to approximate the acceptability judgements of 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 & Wilson (2008)'s approach to create gradient acceptability judgements from the Scholes (1966) experiment, following previous studies (Pierrehumbert, 1994; Coleman & Pierrehumbert, 1997). In Zimmer's study, some words were tested twice, leading to minor variations in response rates, e.g., [tɑtuz] receives either 0.375 and 0.3125, which do not significantly influence the results of the statistical tests below.

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.

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

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

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



(a) Exception-Filtering

(b) HW

Figure 6: Scatterplot 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. In Zimmer's experiment, certain words were used twice, and response rates for both instances were plotted here.

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 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 judgements. The predicted scores of the Exception-Filtering learner are concentrated at 0 and 1, while the exp(−harmony) is on a continuum.[25]

Both regression models reject the null hypothesis that the predicted judgements have no effect on the proportion of "yes" responses (Exception-Filtering:

[25] 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 $[0, +\inf)$ is $(0, 1]$. This value should not be mistaken as probability, despite their similar ranges.

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, such as ɑ...i in [tɑtiz], which can mislead the HW learner to assign relatively higher probabilities than harmonic sequences such as e...e in [pemez].

To summarise, the Exception-Filtering learner trained using a 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 judgement and the approximated acceptability judgement of nonce words from previous behavioural experiment. This result indicates the capability of the Exception-Filtering model in modelling phonotactic patterns with exceptions.

8. Discussion

To summarise the case studies, in terms of interpretability and scalability, 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 judgements 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 (Pinker & Prince, 1988; Zuraw, 2000; Zuraw *et al.*, 2021). Within this framework, extragrammatical factors are conceptualised as originating from two main sources: performance-related and lexicon-related variables. Performance-related variables include individual differences, auditory illusions (Kahng & Durvasula, 2023), and task effects in general (Armstrong *et al.*, 1983; Gorman, 2013). Lexicon-related variables include lexical information such as lexical similarity (Bailey & Hahn, 2001, 2005; Avcu *et al.*, 2023), frequency (Frisch *et al.*, 2000; Ernestus & Baayen, 2003), etc.

In the current study, in tandem with the learnt grammar, extragrammatical factors contribute to acceptability judgements in behavioural experiments. For example, previous studies have shown that lexical similarity and frequency are significant predictors of acceptability judgements (Bailey & Hahn, 2001, 2005; Frisch *et al.*, 2000). Performance-related variables, such as individual differences and task effects, can also influence acceptability judgements. Therefore, a comprehensive evaluation of a learnt grammar against acceptability judgements should take these factors into account. In future research, this evaluation could be carried out 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.

.

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 & Gallagher (2018). Segmental representations may overlook subsegmental generalisations—underrepresented segmental two-factors in the input data can exhibit high frequency in feature-based generalisations. 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. As Hayes & Wilson (2008: 401) demonstrated, a feature-based model outperforms a segment-based model in their English case study.

It is feasible to integrate feature-based representations into the current approach using the generality heuristics in Hayes & 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. 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

	F	*G*
C	+	+
V	+	-

Table 21: Simplified feature system.

lower than the current threshold. For example, a constraint such as $*[+G]\ [+G]$ would imply more specific two-factors such as $*[+G]\ [+F, +G]$, $*[+F, +G]\ [+G]$, 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 & 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, especially the benchmark HW learner, which learns a Maximum Entropy Grammar (Goldwater & 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 by maximising the likelihood of observed input data, including 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 & 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). As a result, researchers usually manually remove the strings considered lexical exceptions from the training data prior to simulations, such as in the English case study of Hayes & Wilson (2008).

This issue has motivated several interesting proposals to handle exceptions within the HW learner. Hayes & Wilson (2008: 386) added a Gaussian prior to prevent overfitting by adjusting the standard deviation σ of the Gaussian distribution for constraint weights. Although this method proves effective for certain datasets based on their specific noise distribution, it still assigns nonzero, albeit low, probabilities to lexical exceptions.

Another strategy is to include lexically specific constraints in the hypothesis space to handle lexical exceptions (Pater, 2000; Linzen *et al.*, 2013; Moore-Cantwell & Pater, 2016; Hughto *et al.*, 2019; O'Hara, 2020). Lexically specific constraints such as $*\mathrm{sf}_i$ 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. Such computational complexity not only restricts our capacity to test the proposal adequately, but also raises questions about its plausibility in child language acquisition.

Both proposals above handle the exception-related overfitting problem through the incorporation of a regularisation 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 anomalies 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 in an accuracy schedule. 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.[26] 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 *gain* as per Della Pietra *et al.* (1997) and Berent *et al.* (2012), and the Tolerance Principle as per Yang (2016). Similar to $\theta_{\max}$ in the accuracy schedule, gain is set at a specific threshold—the higher the gain, the more statistical support is required for a constraint to be added to the hypothesis grammar (Gouskova & Gallagher 2020: 5). 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 involves some nontrivial adjustments, especially deriving a probabilistic distribution from categorical grammars.

The Tolerance Principle proposes that a rule will be generalised if the number of exceptions does not exceed the number of words in the category N divided by the natural log of N ($N/\ln N$). This threshold is set *a priori* for each N before the learner is exposed to the training data, rather than induced as in the current proposal. Although 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 nontrivial adjustment in defining the scope of phonotactic constraint.

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

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

learning model. This has led to significant simplifications in the proposed learning model. Therefore, the 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 & Breiss, 2021). A 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. Second, the learnt grammar in Polish shows a viable approach to interpret SSP-defying onsets in the context of lexical exceptions (Jarosz, 2017). Third, this study prespecifies tiers for the hypothesis space during phonotactic learning. In the future, it would be beneficial to integrate an automatic tier induction algorithm based on the principles proposed in previous studies (Jardine & Heinz, 2016; Gouskova & Gallagher, 2020). Another promising direction is to extend the current approach to the hypothesis space defined by other formal languages (Jäger & Rogers, 2012).

Last but not least, while phonotactic learning facilitates the learning of morphophonological alternations, it cannot independently motivate alternation learning, as shown in experimental studies (Pater & Tessier, 2006; Chong, 2021). Given this evidence, a future direction is to model phonotactic and alternation learning as distinct but interconnected components. The phonotactic model proposed in this paper can be used to filter out lexical exceptions that interfere with alternation learning. For example, in Turkish round harmony, after the rounded stem vowel [ø], the high front vowel /i/ in the suffixes typically changes to round [y]. However, in noisy real-world data, unrounded [i] can exceptionally surface after [ø]. The phonotactic model proposed in the current study can be used to filter out illicit sequences such as [ø…i] during alternation learning, allowing feature-based generalisations such as /i/ → [+round]/[+round].

9. 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. Moreover, while the current study primarily focusses on the learning of categorical grammars, it lays the groundwork for integrating learnt grammars with extragrammatical factors to model behavioural data, and marks initial steps in reassessing the ability of categorical grammars in approximating human judgements.

Competing interests. The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments. I thank Colin Wilson, Adam Jardine, Bruce Tesar, Yang Wang, Adam McCollum, Jeff Heinz, Tim Hunter, Bruce Hayes, Gaja Jarosz, Caleb Belth, Jon Rawski, anonymous reviewers, audience at AMP 2022, UCI QuantLang Lab, and MIT Exp/Comp group for their help and suggestions. All mistakes are mine.

Data availability statement. Data and code can be found in https://github.com/hutengdai/exception-filtering-phonotactic-learner.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Albright, Adam (2007). Natural classes are not enough: biased generalization in novel onset clusters. In *15th manchester phonology meeting, manchester, uk*. 24–26.

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

Albright, Adam & Bruce Hayes (2002). Modeling english past tense intuitions with minimal generalization. In *Proceedings of the acl-02 workshop on morphological and phonological learning*. 58–69.

Albright, Adam & Bruce Hayes (2003). Rules vs. analogy in english past tenses: a computational/experimental study. *Cognition* **90**. 119–161.

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

Altan, Aslı, Utku Kaya & Annette Hohenberger (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*.

Alves, Fernando C (2023). Categorical versus gradient grammar in phonotactics. *Language and Linguistics Compass* **17**. e12501.

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

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

Armstrong, Sharon Lee, Lila R Gleitman & Henry Gleitman (1983). What some concepts might not be. *Cognition* **13**. 263–308.

Avcu, Enes, Olivia Newman, Seppo P Ahlfors & David W Gow Jr (2023). Neural evidence suggests phonological acceptability judgments reflect similarity, not constraint evaluation. *Cognition* **230**. 105322.

Baayen, R Harald, Richard Piepenbrock & Leon Gulikers (1995). The celex lexical database (release 2). *Distributed by the linguistic data consortium, University of Pennsylvania* .

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

Bailey, Todd M & Ulrike Hahn (2005). Phoneme similarity and confusability. *Journal of memory and language* **52**. 339–362.

Berent, Iris, Colin Wilson, Gary F Marcus & Douglas K Bemis (2012). On the role of variables in phonology: remarks on hayes and wilson 2008. *Linguistic inquiry* **43**. 97–119.

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

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

Chomsky, Noam (1965). *Aspects of the theory of syntax*. MIT press.

Chomsky, Noam & Morris Halle (1965). Some controversial questions in phonological theory. *Journal of linguistics* **1**. 97–138.

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

Clark, Alexander & Shalom Lappin (2009). Another look at indirect negative evidence. In *Proceedings of the eacl 2009 workshop on cognitive aspects of computational language acquisition*. 26–33.

Clark, Alexander & Shalom Lappin (2010). *Linguistic nativism and the poverty of the stimulus*. John Wiley & Sons.

Clements, George N. (1990). The role of the sonority cycle in core syllabification. In John Kingston & Mary E. Beckman (eds.) *Papers in laboratory phonology i: between the grammar and physics of speech*. Cambridge University Press, 283–333.

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

Coleman, John & Janet Pierrehumbert (1997). Stochastic phonological grammars and acceptability. In *Computational phonology: third meeting of the ACL special interest group in computational phonology*.

Dai, Huteng, Connor Mayer & Richard Futrell (2023). Rethinking representations: a log-bilinear model of phonotactics. *Proceedings of the Society for Computation in Linguistics* **6**. 259–268.

Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann (2011). Explaining sonority projection effects. *Phonology* **28**. 197–234.

Davis, Stuart & Michael Hammond (1995). On the status of onglides in american english. *Phonology* **12**. 159–182.

Della Pietra, Stephen, Vincent Della Pietra & John Lafferty (1997). Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence* **19**. 380–393.

Dillon, Brian & Matthew W. Wagers (2021). *Approaching gradience in acceptability with the tools of signal detection theory*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 62–96.

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

Eisner, Jason (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*. 313–320.

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

Ernestus, Mirjam Theresia Constantia & R Harald Baayen (2003). Predicting the unpredictable: interpreting neutralized segments in dutch. *Language* **79**. 5–38.

Frisch, Stefan A, Nathan R Large & David B Pisoni (2000). Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of memory and language* **42**. 481–496.

Frisch, Stefan A, Nathan R Large, Bushra Zawaydeh, David B Pisoni *et al.* (2001). Emergent phonotactic generalizations in english and arabic. *Typological studies in Language* **45**. 159–180.

Frisch, Stefan A, Janet B Pierrehumbert & Michael B Broe (2004). Similarity avoidance and the ocp. *Natural language & linguistic theory* **22**. 179–228.

Frisch, Stefan A & Bushra Adnan Zawaydeh (2001). The psychological reality of ocp-place in arabic. *Language* . 91–106.

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

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

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

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

Göksel, Aslı & Celia Kerslake (2004). *Turkish: a comprehensive grammar*. Routledge.

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

Goldsmith, John (1976). *Autosegmental phonology*. PhD dissertation, MIT Press London.

Goldwater, Sharon & Mark Johnson (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.A. & W.H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* **49**. 732–764.

Gorman, Kyle (2013). *Generative phonotactics*. PhD dissertation, University of Pennsylvania.

Gorman, Kyle (2016). Pynini: a python library for weighted finite-state grammar compilation. In *Proceedings of the sigfsm workshop on statistical nlp and weighted automata*. 75–80.

Gouskova, Maria & Gillian Gallagher (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory* . 1–40.

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

Hale, John & Paul Smolensky (2006). Harmonic grammars and harmonic parsers for formal languages. In Paul Smolensky & Géraldine Legendre (eds.) *The harmonic mind, volume 1: from neural computation to optimality-theoretic grammar*. MIT Press, 393–416.

Hale, Mark & Charles Reiss (2008). *The phonological enterprise*. OUP Oxford.

Haman, Ewa, Bartłomiej Etenkowski, Magdalena Łuniewska, Joanna Szwabe, Ewa Dabrowska, Marta Szreder & Marek Łaziński (2011). Polish cds corpus. Available from http://childes.psy.cmu.edu.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman & Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

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

Hayes, Bruce & Zsuzsa Cziráky Londe (2006). Stochastic phonological knowledge: the case of hungarian vowel harmony. *Phonology* **23**. 59–104.

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

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

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

Heinz, Jeffrey, Gregory M Kobele & Jason Riggle (2009). Evaluating the complexity of optimality theory. *Linguistic Inquiry* **40**. 277–288.

Heinz, Jeffrey, Chetan Rawal & Herbert G Tanner (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*. Association for Computational Linguistics, 58–64.

Hohenberger, Annette, Aslı Altan, Utku Kaya, Özgün Köksal Tuncer & Enes Avcu (2016). Sensitivity of turkish infants to vowel harmony: preference shift from familiarity to novelty. In F. Nihan Ketrez & Belma Haznedar (eds.) *The acquisition of turkish in childhood*. John Benjamins Publishing Company, 29–56.

Hughto, Coral, Andrew Lamont, Brandon Prickett & Gaja Jarosz (2019). Learning exceptionality and variation with lexically scaled maxent. In *Proceedings of the society for computation in linguistics (scil) 2019*. 91–101.

Hyman, Larry M (1975). *Phonology: theory and analysis*. Holt, Rinehart & Winston.

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

Inkelas, Sharon, Aylin Küntay, C. Orhan Orgun & Ronald Sprouse (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, Ray (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford University Press.

Jäger, Gerhard & James Rogers (2012). Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**. 1956–1970.

Jardine, Adam (2016). Learning tiers for long-distance phonotactics. In *Proceedings of the 6th conference on generative approaches to language acquisition north america (galana 2015)*. 60–72.

Jardine, Adam & Jeffrey Heinz (2016). Learning tier-based strictly 2-local languages. *Transactions of the Association for Computational Linguistics* **4**. 87–98.

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

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

Jarosz, Gaja, Shira Calamaro & Jason Zentz (2017). Input frequency and the acquisition of syllable structure in polish. *Language acquisition* **24**. 361–399.

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

Jusczyk, Peter W & Richard N Aslin (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology* **29**. 1–23.

Jusczyk, Peter W, Angela D Friederici, Jeanine MI Wessels, Vigdis Y Svenkerud & Ann Marie Jusczyk (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of memory and language* **32**. 402–420.

Jusczyk, Peter W, Paul A Luce & Jan Charles-Luce (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of memory and Language* **33**. 630–645.

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

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

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

Kawahara, Shigeto & Canaan Breiss (2021). Exploring the nature of cumulativity in sound symbolism: experimental studies of pokémonastics with english speakers. *Laboratory Phonology* **12**.

Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika* **30**. 81–93.

Kostyszyn, Kalina & Jeffrey Heinz (2022). Categorical account of gradient acceptability of word-initial polish onsets. In *Proceedings of the annual meetings on phonology*, volume 9.

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

Lau, Jey Han, Alexander Clark & Shalom Lappin (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science* **41**. 1202–1241.

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

Lewis, Geoffrey L. (2001). *Turkish grammar*. 2nd edition. Oxford University Press.

Lignos, Constantine & Kyle Gorman (2012). Revisiting frequency and storage in morphological processing. In *Proceedings from the annual meeting of the chicago linguistic society*, volume 48. Citeseer, 447–461.

Linzen, Tal, Sofya Kasyanenko & Maria Gouskova (2013). Lexical and phonological variation in russian prepositions. *Phonology* **30**. 453–515.

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

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

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

Mayer, Connor, Adam McCollum & Gülnar Eziz (2022). Issues in uyghur phonology. *Language and Linguistics Compass* **16**. e12478.

McMullin, Kevin & Gunnar Ólafur Hansson (2019). Inductive learning of locality relations in segmental phonology. *Laboratory Phonology* **10**.

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

Mohri, Mehryar, Afshin Rostamizadeh & Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.

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

Moreton, Elliott, Joe Pater & Katya Pertsova (2017). Phonological concept learning. *Cognitive science* **41**. 4–69.

Nevins, Andrew & Bert Vaux (2003). Metalinguistic, shmetalinguistic: the phonology of shmreduplication. In *Proceedings from the annual meeting of the chicago linguistic society*, volume 39. Chicago Linguistic Society, 702–721.

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

Osherson, Daniel, Michael Stob & Scott Weinstein (1986). *Systems that learn: an introduction to learning theory*. MIT press.

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

Pater, Joe & Anne-Michelle Tessier (2006). L1 phonotactic knowledge and the l2 acquisition of alternations. In R. Slabakova, S. Montrul & P. Prévost (eds.) *Inquiries in linguistic development: studies in honor of lydia white*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 115–131.

Pearl, Lisa & Jeffrey Lidz (2009). When domain-general learning fails and when it succeeds: identifying the contribution of domain specificity. *Language Learning and Development* **5**. 235–265.

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, Janet (1993). Dissimilarity in the arabic verbal roots. In *Proceedings of nels*, volume 23. University of Massachusetts Amherst, 367–381.

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

Pierrehumbert, Janet (2001). Stochastic phonology. *Glot international* **5**. 195–207.

Pinker, Steven & Alan Prince (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* **28**. 73–193.

Prince, Alan & Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. Blackwell.

Prince, Alan & Bruce Tesar (2004). Learning phonotactic distributions. In *Constraints in phonological acquisition*. Cambridge University Press Cambridge, 245–291.

Rawski, Jonathan (2021). *Structure and learning in natural language*. PhD dissertation, State University of New York at Stony Brook.

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

Richtsmeier, Peter T (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology* .

Rose, Sharon & Lisa King (2007). Speech error elicitation and co-occurrence restrictions in two ethiopian semitic languages. *Language and Speech* **50**. 451–504.

Rubach, Jerzy & Geert Booij (1990). Syllable structure assignment in polish. *Phonology* **7**. 121–158.

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

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

Shattuck-Hufnagel, Stefanie (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**. 72–101.

Sundara, Megha, ZL Zhou, Canaan Breiss, Hironori Katsuda & Jeremy Steffman (2022). Infants' developing sensitivity to native language phonotactics: a meta-analysis. *Cognition* **221**. 104993.

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

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

Weide, Robert *et al.* (1998). The Carnegie Mellon pronouncing dictionary. *Release 06, wwwcscmuedu* .

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

Wilson, Colin & Gillian Gallagher (2018). Accidental gaps and surface-based phonotactic learning: a case study of south bolivian quechua. *Linguistic Inquiry* **49**. 610–623.

Wilson, Colin & Marieke Obdeyn (2009). Simplifying subsidiary theory: statistical evidence from arabic, muna, shona, and wargamay. Ms. Johns Hopkins University.

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

Wu, Katherine & Jeffrey Heinz (2023). String extension learning despite noisy intrusions. In *International conference on grammatical inference*. PMLR, 80–95.

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

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

Zuraw, Kie, Isabelle Lin, Meng Yang & Sharon Peperkamp (2021). Competition between whole-word and decomposed representations of english prefixed words. *Morphology* **31**. 201–237.

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

Zydorowicz, Paulina & Paula Orzechowska (2017). The study of polish phonotactics: measures of phonotactic preferability. *Studies in Polish Linguistics* **12**.

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. 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 responses indicating "no preference" were excluded.

Experiment 1				Experiment 2			
Harmonic		Disharmonic		Harmonic		Disharmonic	
temez	19	temɑz	3	pemez	30	pemɑz	2
teriz	23	terɯz	0	teriz	28	terɯz	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, adapted from Gorman (2013: §3.2.2).

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

Table 23: Effects of roundness harmony on Zimmer (1969)'s wordlikeness experiment, adapted from Gorman (2013: §3.2.2).

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	d͡ʐ	18	m n	8	d b	4	m d͡z d	2	f ʂ t͡ʂ	1
k	2,052	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	t͡s 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	t͡s f	8	p x w	3	t͡s 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	d͡z b	8	r j	3	f t͡s	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 t͡s	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 d͡z	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	d͡z	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	d͡z 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 t͡s	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
t͡s	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	t͡s 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	d͡ʐ d͡ʐ	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 d͡z	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