PHONOLOGICAL LEARNING IN THE PRESENCE OF LEXICAL EXCEPTIONS

By

HUTENG DAI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Linguistics

Written under the direction of

Adam Jardine

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May 2024

All generalizations are false, including this one.

—Mark Twain

ABSTRACT OF THE DISSERTATION

Phonological learning in the presence of lexical exceptions

by HUTENG DAI

Dissertation Director: Adam Jardine

In this dissertation, I establish a research program that uses computational modeling as a testbed for theories of phonological learning. This dissertation focuses on a fundamental question: how do children acquire sound patterns from noisy, real-world data, especially in the presence of lexical exceptions that defy regular patterns? For instance, Turkish infants tune into vowel harmony patterns as early as six months, despite lexical exceptions from disharmonic loanwords. This dissertation demonstrates that phonological learning is facilitated by two essential elements: (1) a restrictive hypothesis space defined by formal language theory and (2) an exception-filtering mechanism. I developed a learning model that harnesses the discrete nature of categorical grammars to filter out lexical exceptions based on statistical criteria adapted from probabilistic models. This hybrid model learns interpretable grammars that approximate acceptability judgments in behavioral experiments, demonstrating robust performance across various real-world corpora from English, Polish, and Turkish. Moreover, the dissertation integrates the proposed phonotactic model into learning morpho-phonological alternations. This approach is not only competitive on real-world corpora but also substantiated by experimental evidence.

ACKNOWLEDGMENTS

I'd like to express my deepest gratitude to colleagues, family, and friends who shaped both my career and personal growth. This list is merely a small subset of many people I wish to thank.

My advisor, Adam Jardine, introduced me to the realm of computational linguistics. He kept me grounded during moments of unwarranted confidence, offered unwavering support through times of doubt, and celebrated each small step I took towards the tipping point of scientific discovery. He is the best mentor who embodies the character I aspire to develop in myself.

Bruce Tesar ignited my interest in computational learning theory and taught me many life lessons, such as "keep marching forward even when things go horribly wrong" and "find a project that is big enough to be interesting but small enough to be doable". It's been an honor to be your student and to work with you.

Adam McCollum advised my second qualifying paper during the COVID-19 pandemic. Without his support during that period, I wouldn't be here writing my dissertation. I'm grateful for the time he spent commenting on my many drafts and am still amazed by his humor and professionalism, qualities that I wish I had more time to acquire.

Richard Futrell and I met at the 2019 LSA Summer Institute and started our long-term collaboration. I often tell others that he is the smartest person I've ever met. Thanks to Richard, I learned to be a divergent thinker and discovered the connection between seemingly incompatible frameworks and theories.

I owe thanks to Colin Wilson for his aid in developing an efficient algorithm for expected frequencies. My work stands on the shoulders of his and Bruce Hayes' seminal contributions to phonotactic learning. Additionally, I am thankful to Bruce Hayes for sharing the English training data, to Gaja Jarosz for the Polish data, and to Caleb Belth for both his valuable feedback and the Turkish data he provided.

Throughout my journey as a language scientist, I've had the privilege of being guided by

an extraordinary group of mentors. Connor Mayer, Jeff Heinz, Andrew Lamont, Kristen Syrett, Yimei Xiang, Troy Messick, Simon Charlow, Eric Baković, Ryan Bennett, Matt Gordon, Adam Albright, and Zihe Li have each, in their own unique way, contributed to my development as a scholar. Their wisdom and support have been pillars upon which I've built my research and confidence. It is the opportunity to meet exceptional individuals like these colleagues that makes an academic career so fulfilling.

Life is needlessly complex at times, but I could always find solid ground by connecting with my family. It's not easy to be the parents of an international PhD student, yet my parents, Dai Ben and Liao Chenghong, have embraced the challenge with exceptional grace. Their belief in my vision has been a constant source of strength and motivation.

The camaraderie and support of friends have enriched my PhD experience immeasurably. I will name a few here: Yang Wang, Jill Harper, Ryan Rhodes, Jennifer Kuo, Tajudeen Yacoubou Mamadou, Jon Rawski, Scott Nelson, Chaoyi Chen, Sreekar Raghotham, Marjorie Leduc, Scott James, Brian Pinsky, Gerry Avelino, Hyunjung Joo, Seoyoung Kim, Vinny Czarnecki, and Jinyoung Jo. The bonds formed during this time have been a source of joy, inspiration, and resilience.

The creation of a dissertation is a collaborative effort between the writer and the reader. To the colleague who is reading this work, I extend my heartfelt thanks. If you happen to be in the valley of despair, questioning the worth of this journey, I'd like to offer some insights that have crystallized from my recent reflections: although a significant portion of scientific work, including this dissertation, may eventually fade into obscurity, our seemingly needless efforts as scientists play a crucial role in the incremental progress of human understanding. Let's keep stretching our arms and reaching farther than we did yesterday.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This dissertation tackles a central question in linguistics: how do children acquire their languages, particularly phonological knowledge, from the noisy real-world data surrounding them? This dissertation posits that effective phonological learning is underpinned by two elements: (1) a *restrictive hypothesis space*, as defined by formal language theory (Heinz, 2007; Jardine and Heinz, 2016; Dai and Futrell, 2021), and (2) an *exception-filtering mechanism*. In particular, I introduce a hybrid model that combines the discrete nature of traditional categorical grammars with statistical criteria to detect and filter out lexical exceptions in phonological learning.

1.1 Phonological Learning in the Real World

Let's begin with the phonological knowledge in our textbooks. There are two main categories: *phonotactics*, which determines the permissible (or well-formed) combinations of sounds. For example, English speakers disfavor *bn and *sf sequences. Phonotactic knowledge is often tested using novel form judgments, also known as *blick* test: speakers may judge **bnick* and **sfid* as unacceptable (Scholes, 1966), while accepting novel forms such as *blick*.

The second category, *alternations*, involves the different realizations of underlying representations (URs) based on morphophonological context; for instance, the English plural manifests in various forms depending on the preceding phonological environment. Alternation knowledge can be evaluated through the well-known *wug test* (Berko, 1958), where, for example, the English plural is articulated as voiced [z] in [wʌgz].

Textbook examples typically depict a polished, exception-free overview of phonological knowledge. However, numerous exceptions, or lexical exceptions, contravene these phonologi-

cal generalizations in the real world, as Edward Sapir famously noted, "All grammars leak." The word *sphere* in the English lexicon, for example, is an exception to the onset phonotactics *sf. Table Table 1.1 lists other lexical exceptions in English that defy onset phonotactics confirmed through behavioral experiments (Scholes, 1966). Many of these exceptions are borrowed from other languages:

onset	**exception**	**origin**
*sf	*sphere*	Greek
*zl	*zloty*	Polish
*bw	*Bois*	French
*sr	*Sri*	Sanskrit
*ʃl	*schlep*	Yiddish

Table 1.1: English onset phonotactics, lexical exceptions, and origins.

These lexical exceptions are widespread and sometimes isomorphic across both phonotactics and alternations (see discussion in chapter 6). For instance, the standard pattern of Turkish back vowel harmony typically penalizes nonlocal occurrences of vowels with differing backness features. The plural suffix /-lAr/ is realized as [-ler] following a front vowel, e.g., [i**p**-l**e**r] 'rope', [k**ø**y-l**e**r] 'village', and as [-lar] following a back vowel, e.g., [k**ɯ**z-l**a**r] 'girl', [p**u**l-l**a**r] 'stamp'. Nevertheless, numerous disharmonic loanwords resulting from language contact with Arabic, French, or English do not undergo vowel harmony, as shown in Table Table 1.2.

stem-plural	gloss	**violated constraints**	**origin**
sinj**a**l-l**e**r	signal	*back…front (*ɑ…e)	English
dikk**a**t-l**e**r	attention	*back…front (*ɑ…e)	Arabic
protok**o**l-l**e**r	protocol	*back…front (*o…e)	French

Table 1.2: Lexical exceptions of Tukrish vowel harmony and origins

Despite the prevalence of lexical exceptions, children learn phonological patterns remarkably early in life. English infants, for example, become attuned to onset phonotactics between 8-10 months, and Turkish infants demonstrate awareness of vowel harmony's nonlocal phono-

tactics by 6 months (Altan et al., 2016; Sundara et al., 2022). The perceptual sensitivity of phonological alternations develops as early as 12 months (White et al., 2008; Skoruppa et al., 2013). How do children achieve such impressive learning outcomes amid noisy input?[1]

1.2 Why Computational Modeling?

This dissertation employs computational modeling to address the overarching question linking theoretical phonology, acquisition, and cognitive science. Computational approaches require researchers to precisely define theoretical concepts algorithmically, facilitating the development of a mathematically well-defined, computational theory of phonological learning. Moreover, computational modeling based on realistic corpora provides a rigorous testing ground for linguistic theories. Figure Figure 1.1 illustrates the comprehensive research program: children's linguistic environments can be simulated with realistic data, and computational models are trained to learn grammars that replicate human linguistic judgments in experimental settings, such as the "blick" and "wug" tests. Such methodologies can probe into the types of *information* and *structures* that children must attend to or disregard during phonological learning.

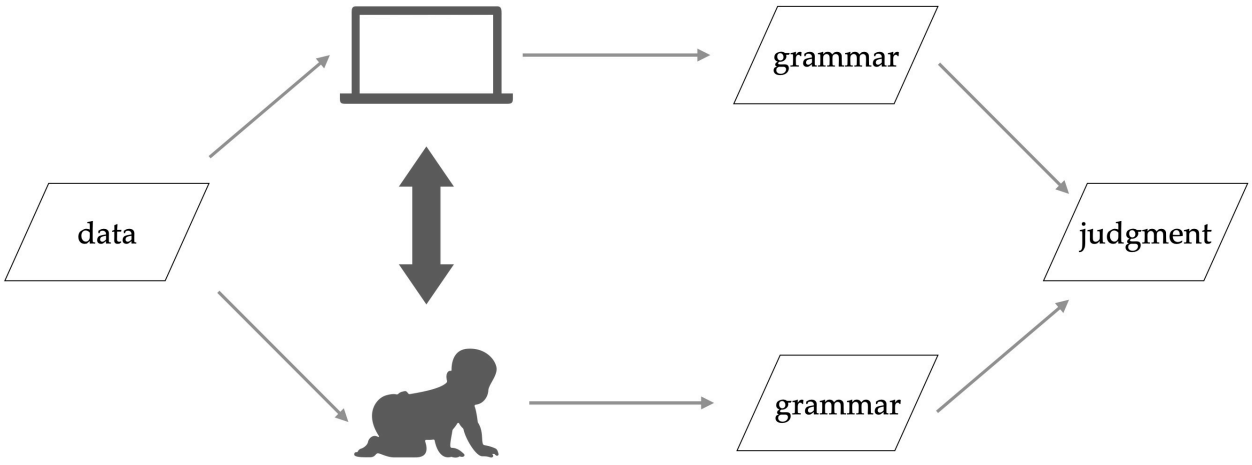


Figure 1.1: The overarching research program: a computational theory of learning; the machine icon represents the computational model.

[1]The present study operates under the assumption that children's linguistic input includes lexical exceptions, albeit not as extensively as the input encountered by second language learners.

Furthermore, the interplay between computational modeling and language acquisition is inherently reciprocal. Insights into language acquisition inform the development of advanced machine learning algorithms by illuminating fundamental aspects of learning—how any learning devices/machines can learn and achieve human linguistic cognition from noisy real-world data. This parallelism is due to similarities in input and output between child language acquisition and machine learning, aligning their learning objectives at a computational level (Marr, 1982; Tesar and Smolensky, 2000). Additionally, the cognitively inspired computational models provide practical tools for phonological analysis, especially low-resource corpora.

This work is situated at a time when domain-general Large Language Models (LLMs), such as the Transformer architecture (Vaswani et al., 2017), are gaining prominence. Characterized by their dependence on vast datasets, significant computational power, and uninterpretable parameters, these models have shown remarkable ability to mimic human linguistic behaviors. Naturally, this engineering leap prompts a critical inquiry in this study: what is the value and contribution of constructing interpretable, domain-specific computational models for language learning?

Besides interpretable models that are necessary for any scientific theories, the contributions of this dissertation lies in its endeavor to model the relatively modest, unstructured, and noisy linguistic input that children typically encounter. The extensive and often pre-processed datasets that underpin the training of contemporary LLMs do not adequately mirror the realistic learning environment of a child. Consequently, the success of these models under such artificially enhanced conditions does not directly advance our understanding of human language acquisition. In chapter 7, the current study further extends its scope by comparing the learning outcomes derived from my proposals with those obtained from state-of-the-art Transformer-based seq2seq architectures. Through this comparison, the dissertation aims to highlight the distinctive insights and contributions that domain-specific, interpretable models bring to our comprehension of human language acquisition beyond engineering advances.

1.3 Main proposals and findings

Following the computational methodology introduced above, the current study proposes models of phonotactic and alternation learning. Phonological learning is boiled down to two essential ingredients: (1) a restrictive hypothesis space defined by Formal Language Theory, namely Tier-based Strictly Local Languages and functions, and (2) an exception-filtering mechanism that iteratively filters out lexical exceptions from the input data. In particular, the current study shows the power of "categorical grammars + statistical criteria" approach in handling lexical exceptions (Yang, 2016; Durvasula, 2020; Kostyszyn and Heinz, 2022). This hybrid model learns interpretable grammars that approximate acceptability judgments in behavioral experiments, demonstrating robust performance across diverse real-world corpora from English, Polish, and Turkish.

Moreover, this dissertation integrates the proposed phonotactic learning model into the learning of morphophonological alternations. The learned phonotactic grammar is utilized to filter out lexical exceptions in alternation learning. This is not only practically convenient, but also grounded in empirical evidence that phonotactic learning precedes and facilitates alternation learning.

1.4 Roadmap

This dissertation unfolds as follows: chapter 2 lays out the foundational background of the entire dissertation; chapter 3 introduces the core proposal of phonotactic learning; chapter 4 evaluates the phonotactic learning model proposed in chapter 3; chapter 5 discusses the open questions related to phonotactic learning; chapter 6 extends the phonotactic learning model to include alternation learning and presents the two-stage phonotactic-alternation learning model; chapter 7 evaluates the two-stage phonotactic-alternation learning model in realistic corpora.

The contributions of this dissertation do not signify the end of the research program. Rather, they represent the incremental progress in advancing the understanding of human language ac-

quisition and cognition, laying the foundation for future research that will refine the proposals presented herein. The chapter 8 delves into new questions and topics arising from this study and outlines the directions for future work. The chapter 9 provides a summary of the entire dissertation.

CHAPTER 2

BACKGROUND

This chapter outlines the essential concepts, underlying assumptions, and relevant evidence involved in the current proposal, starting with the assumption about phonological grammar and the architecture of phonological processing.

2.1 The Role of Grammar in Phonological Processing and Learning

This study assumes three interconnected components involved in the computation from a word to its acceptability: grammar, lexicon, and performance. The relationship between these components is visualized in Figure 2.1. The acceptability judgment is influenced by both competence and performance factors. For example, [sfid] is not attested in the lexicon. When the grammar penalizes *sf sequence, the grammaticality of [sfid] is 0 ("ungrammatical"). Together, the lexicon and grammar form the *competence*, representing the internalized knowledge, and they predict a relatively low acceptability of [sfid], assuming insignificant influence of other performance factors.

Behavioral experiments employing nonce words aim to mitigate lexical influence, thereby focusing on grammaticality judgments. Nonetheless, participants' sensitivity to lexical similarity with existing items stored in the lexicon, aka. neighborhood density (Vitevitch and Luce, 1998), may still affect their acceptability judgments. Phonological learning discussed in the current study focuses on the the acquisition of grammars from primary linguistic data.[1]

[1]The Fragment Grammar in O'Donnell et al. (2009) provides an interesting perspective to the division of Lexicon and Grammar based on Bayesian modeling.

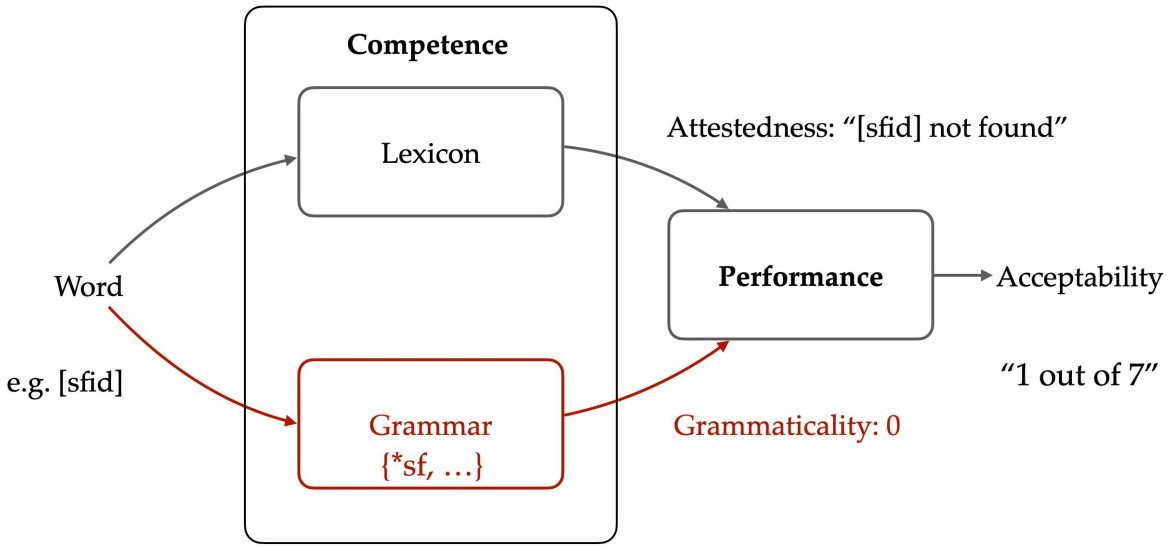


Figure 2.1: The processing of human acceptability judgments. The acceptability is influenced by lexical, grammatical, and various performance factors.

Although the terms *grammaticality* (or well-formedness) and *acceptability*. have frequently been conflated in previous research (Albright, 2009), the current study refers to "grammaticality" only as the predicted score by the hypothesis grammar. In this context, acceptability refers to the judgments made by native speakers on real-world performance, which can be influenced by both grammar and extragrammatical factors, such as processing difficulty, lexical frequency, and neighborhood density (Schütze, 1996, see detailed discussion in chapter 5). In contrast, grammaticality is the output of the abstract, internalized knowledge represented by the grammar, such as phonotactic constraints, 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 and 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 could 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 and Durvasula 2023:3).

2.2 Gradient Judgments and Categorical Grammar

The current study assumes that the grammaticality of sound sequences, categorical or probabilistic, is *reflected* in acceptability judgments, and a successful grammar should exhibit a robust correlation between predicted grammaticality and acceptability judgments to allow "direct investigation" of linguistic competence (Lau et al., 2017).

This assumption, however, has increasingly led to a conflation of grammaticality and acceptability in the literature. The term *grammar* has been frequently conflated with any system that can predict human judgment (Chomsky, 1965a). A gradient grammar that directly generates gradient judgments in the behavioral data becomes appealing, and Hayes (2000)'s argument is the most representative:

> I conclude that the proposed attribution of gradient well-formedness judgments to performance mechanisms would be uninsightful. Whatever "performance" mechanisms we adopted would look startlingly like the grammatical mechanisms that account for non-gradient judgments. For this reason, I will assume that the competence model itself should generate gradient judgments.
>
> —Hayes (2000) "*Gradient Well-Formedness in Optimality Theory*"

Hayes is skeptical about performance models and instead argued for a gradient "competence model" that can predict gradient judgment, as visualized in Figure 2.2.

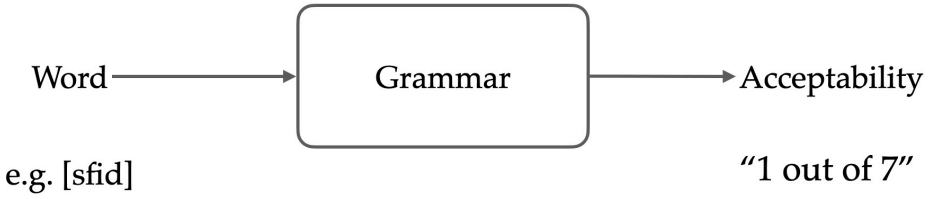


Figure 2.2: A visualization of the gradient "competence model" in Hayes (2000)

However, Hayes (2000)'s assumption that grammaticality equates to acceptability is oversimplified and contradicted by studies on task effects of numeric ratings (Armstrong et al., 1983). In Armstrong et al. (1983), participants provided Likert ratings for inherently categorical concepts such as odd and even numbers. Gradient acceptability collected through numerical rating tasks does not necessitate gradient / probabilistic grammars nor negate categorical grammars (cf. Coleman and Pierrehumbert, 1997; Hayes and Wilson, 2008). His proposal also represents a step towards the conflation of lexical, grammatical, and performance factors in constraint-based frameworks such as Optimality Theory.

Acceptability judgments are commonly collected via rating tasks employing a numeric Likert scale and characterized 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

[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 judgments.

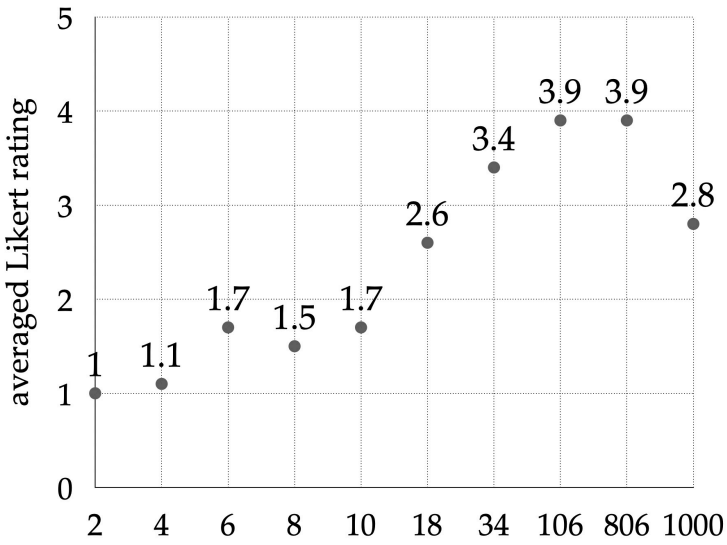


Figure 2.3: Gradient judgments of "How even is number _?" in Armstrong et al. (1983).

incorporate individual variability within a categorical framework (see chapter 5 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), as shown in Figure 2.3.

Another extragrammatical factor at play in the acceptability judgment is traced back to *auditory illusions*, as shown in Kahng and Durvasula (2023).

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 and Wilson, 2008), involve constraints along with continuous weights, assigning a probability continuum across all possible sequences. Analogous to probabilistic grammars, grammaticality in categorical grammars is associated with discrete, often binary values, where 0 signifies ungrammatical sequences, and 1 designates grammatical ones. However, probabilistic grammars with thresholds (Hale and Smolensky, 2006) cannot define infinite languages, as mathematically demonstrated in Alves (2023), and cannot be conflated with categorical grammars.

In the context of phonological learning, choosing between probabilistic and categorical grammars involves appreciating their unique aspects, especially given the lack of definitive evidence favoring either approach. Probabilistic grammars have been noted for their ability to model human sensitivity to frequency information and approximate gradient acceptability judgments (Hayes and Wilson, 2008), while categorical grammars delineate a clear boundary between grammatical words and lexical exceptions (Yang 2016:3). This discrete nature can be used to facilitate phonological learning, as shown in the current study. Moreover, categorical grammars are natural in capturing *overregularizations* in child language acquisition. Children learn discrete regular generalizations (rules or constraints) and overapply them to any novel forms, but they rarely overirregularize patterns in exceptions, which is predicted by learning models based on probabilistic grammars (Yang, 2016; Shi and Emond, 2023).

2.3 Exceptionality

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 1965b: 229; Chomsky and 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* in the input data (drawn from the lexicon) and *grammaticality w.r.t.* the hypothesis grammar. Grammaticality indicates whether phonological representations conform to the hypothesis grammar internalized by the learner. Researchers have used various converging methodologies to approximate the hypothesis grammar, especially statistical generalizations (e.g., observed-to-expected ratio; detailed below) or performance data such as nonce word acceptability (detailed below) and speech errors.[3] For convenience in the dis-

[3]Speech errors elicitation have been used to probe phonotactic constraints (Fromkin, 1973), such as nonlocal

cussion, 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 2.1.

	grammatical	**ungrammatical**
attested	[**br**ɪk]	[*__sf__iɚ]
unattested	[**bl**ɪk]	[*__bn__ɪk]

Table 2.1: The distinction between attestedness and grammaticality (adapted from Hyman, 1975).

This discrepancy between attestedness and grammaticality yields both accidental gaps (grammatical but unattested) and lexical exceptions (attested but ungrammatical), with this dissertation particularly emphasising the latter. For example, although both are nonexistent words, *blick* is grammatical while **bnick* is not, as native speakers uniformly reject **bnick* while accepting *blick*, a classic example of accidental gaps (Chomsky and Halle, 1965; Hayes and Wilson, 2008).[4]

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 and Heinz, 2022). The architecture in Figure 2.1 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

consonant cooccurrences (Rose and King, 2007).

[4]In fact, whether or not *blick* is in the English lexicon is disputable. For example, Dick Blick started an artist supply company over 100 years ago (pc. Adam McCollum; https://www.dickblick.com/about-blick/history/).

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**ʌŋ] (31 "yes"), and even lower than words with unattested onsets such as [**ml**ʌŋ] (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*.[5]

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*, [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 judgments (Scholes, 1966, also see §4.2). Similar examples have been observed in other languages where putative phonotactic restrictions do not extend to loanwords (Gorman 2013:6-7). Thus, this dissertation takes the position that input data drawn from the lexicon can contain ungrammatical lexical exceptions according to the hypothesized phonotactic grammar.

2.4 Rejecting a Nihilistic Perspective on Phonotactics

Hale and Reiss (2008:18) adopted a nihilistic view of phonotactic knowledge, arguing that phonotactics is not part of phonological grammar, as it is computationally inert in morphophonological alternations (Reiss 2017:§6). The benefit of this proposal is that it separates the unproductive phonotactic generalizations from the productive phonological knowledge. The nihilistic view of phonotactics also circumvents the limitations of categorical grammars (Reiss, 2017, 14; "categorical baby"), which have been shown to be ineffective in capturing gradient judgments. See the

[5] [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 §4.2.

above discussion for more details.

However, experimental evidence has shown that infants do acquire morphologically agnostic phonotactics, and the learned phonotactics can facilitate the learning of morphophonological alternations (Jusczyk et al., 1993, 1994; Jusczyk and Aslin, 1995; Archer and Curtin, 2016; Chong, 2021). Gorman (2013:§1) also demonstrates internalization of phonotactic constraints in various domains, such as wordlikeness judgments and loanword adaptation. The current study upholds the concept of categorical grammars, which partly motivated the adoption of the nihilistic view. 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 chapter 5).

2.5 Summary

This chapter has underscored the tension between competence and performance and clarified the nuanced distinctions between acceptability and grammaticality. Computational learning models should correlate the grammaticality scores predicted by the learned grammar with acceptability judgments. The next chapter proposes a learning model that handles lexical exceptions by using an exception-filtering mechanism based on frequency information.

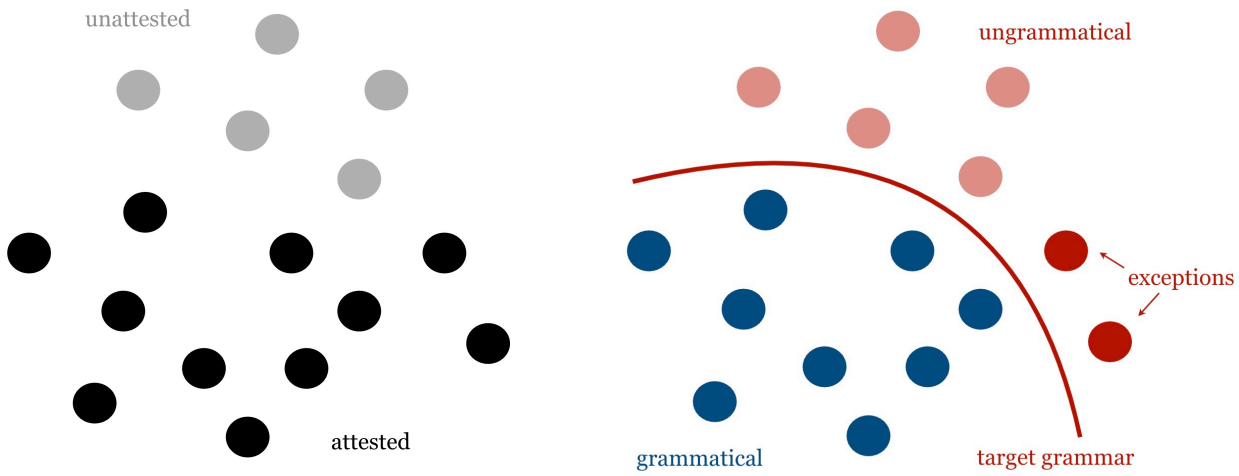
CHAPTER 3

EXCEPTION-FILTERING PHONOTACTIC LEARNER

This chapter proposes a "categorical grammar + exception-filtering" approach to select a hypothesized 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 §3.5.

3.1 Introduction

There exist logically infinite potential sound sequences in any given language, yet only some are considered well-formed by native 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 §4.4 for details). The challenge of phonotactic learning in the presence of lexical exceptions is illustrated in Figure Figure 3.1. Under the positive evidence-only assumption, the learner relies exclusively on unlabelled input data (Marcus, 1993), denoted by the darker dots in the figures; conversely, the lighter dots represent unattested data that are absent from the input. The learning problem is to arrive at a target grammar that can differentiate between grammatical and ungrammatical sequences, represented by the curve in Figure Figure 3.1b.



(a) Potential sound sequences.

(b) Target grammar.

Figure 3.1: The learning problem in the presence of exceptions. Darker dots represent attested data, while light dots indicate unattested onsets; inspired by Mohri et al. (2018:8).

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

Although exceptionality has been a perennial interest in phonology (Wolf, 2011; Moore-Cantwell and Pater, 2016; Mayer et al., 2022)[1], learning models based on categorical grammars capable of handling exceptions 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

[1]The challenge of exception in phonotactic learning is analogous to that of "Type IV" patterns in Moreton et al. (2017), which can be conceptualized 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.

in Gouskova and Gallagher (2020, added emphasis and adapted spelling):

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

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

The current study undertakes a similar endeavour: rooted in formal language theory, it proposes 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 and 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 judgments 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.

3.2 Segment-based Representation

The current proposal adopts segmental representations derived from the input data, a departure from the prespecified feature representations advocated by previous studies (Hayes and Wilson, 2008; Gouskova and Gallagher, 2020). Segmental representations risk misinterpreting accidental gaps as systematic constraints and may overlook sub-segmental generalizations. As Hayes and Wilson (2008:401) demonstrated, a feature-based model outperforms a segment-based model in their English case study. The problem of selecting feature-based constraints is beyond the scope of this study, while chapter 5 demonstrates a promising solution.

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

3.3 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 and Pierrehumbert, 1997; Breiss, 2020; Kawahara and Breiss, 2021), and primarily investigates whether phonotactic learning of categorical grammars is possible in the presence of exception. One possibility to incorporate cumulativity in the future could involve replacing the grammaticality function with the sum of constraint violations (see also chapter 5).

This structure of grammars, while similar, diverges significantly from the cumulative, violable, and ranked grammar in Optimality Theory (OT; Prince and Smolensky, 1993; Prince and Tesar, 2004). The hypothesis grammar in the current proposal is drawn from a highly restrictive hypothesis space.[2] Based on the analytical results of formal language theory (FLT), the current study adopts Tier-based Strictly k-Local (TSL$_k$) languages (Heinz et al., 2011; Jardine and Heinz, 2016; Lambert and Rogers, 2020) as the hypothesis space. In formal language theory, the meanings of "language" deviate from their typical meanings. A language is a set of strings (e.g., sound sequences) that adhere to its associated grammar, which can be mathematically characterized as a set of forbidden or required structures.

k-factors are substrings of length k. A TSL$_k$ grammar consists of all forbidden k-factors on a specific tier, known as TSL$_k$ constraints. The tier, also referred to as a *projection* (Hayes and Wilson, 2008), functions as a targeted subset of the inventory of phonological representations (e.g., segments, consonants, vowels) for constraint evaluation. In the context of local phonotactics, the tier encompasses the full inventory, such as all segments, while in nonlocal phonotactics, it includes only specific segments, such as vowels. For example, as shown in Figure Figure 3.2, 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 lo-

[2]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).

cal constraint *øi on the vowel tier. This concept, although similar, is distinct from the traditional feature-based definition in Autosegmental Phonology where each segment on the projected tier can be linked to multiple sites on the segmental tier (Goldsmith, 1976).

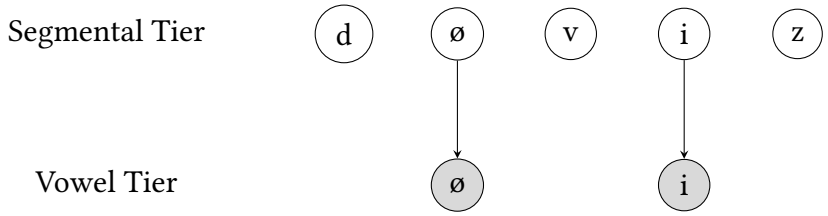


Figure 3.2: Extraction of vowel tier from the Turkish word [døviz] "currency". The vowel tier contains the vowels in this word, disregarding the non-tier consonants.

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 formalized by the function $\texttt{factor}(s, k)$, which generates all k-factors of a string s. For example, $\texttt{factor}(\text{CCV}, 2) = \{\text{CC}, \text{CV}\}$, and $\texttt{factor}(\text{CVC}, 2) = \{\text{CV}, \text{VC}\}$. The grammaticality score of a string s under a grammar G, denoted as $g(s, G)$, is defined as follows:

$$g(s, G) = \begin{cases} 1, & \text{if } \texttt{factor}(s, k) \cap G = \emptyset \\ 0, & \text{if } \texttt{factor}(s, k) \cap G \neq \emptyset \end{cases} \tag{3.1}$$

For example, consider a grammar $G = \{$*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 and 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 and Heinz, 2016; Jardine and McMullin, 2017). This approach has been successfully applied in previous work spanning both probabilistic and categorical approaches (Hayes and Wilson, 2008; Gouskova and Gallagher, 2020; Mayer, 2021; Dai et al., 2023; Heinz, 2007; Jardine and Heinz, 2016).

One of the main challenges of phonotactic learning, as mentioned in Hayes and Wilson (2008:392), is the rapid growth of the hypothesis space with increasing size of k. In response to this challenge, similar to Hayes's solution, 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 dissertation only examines local phonotactics of English and Polish onsets and nonlocal phonotactics of Turkish vowels, the proposed hypothesis space is broadly applicable for suitable domains, extending to phenomena such as nonlocal laryngeal phonotactics in Quechua (Gouskova and Gallagher, 2020), Hungarian vowel harmony (Hayes and Londe, 2006), and Arabic OCP-Place patterning (Frisch and Zawaydeh, 2001; Frisch et al., 2004). To summarize, the learner hypothesizes a noncumulative, inviolable, and unranked categorical TSL_2 grammar, derived from the hypothesis space of TSL_2 languages.

3.4 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 and Heinz (2023).

One approach to address the challenge of exceptions utilizes an *exception-filtering* mecha-

nism to exclude exceptions while learning categorical grammars. Hayes and Wilson (2008:427-428) hypothesized 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 and Lappin 2010:105), the current study achieves this by leveraging *indirect negative evidence* derived from frequency information (Clark and Lappin, 2009; Pearl and Lidz, 2009; Yang, 2016), specifically from type frequency (Pierrehumbert, 2001a; Hayes and Wilson, 2008; Richtsmeier, 2011).[3] 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 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, 2001a; Frisch et al., 2004; Hayes and Wilson, 2008) since Trubetzkoy (1939:Chapter VII).[4] 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

[3]Lexical exceptions might also exhibit unexpectedly high *token frequencies*. For example, in a Wiki corpus of approximately 100 million words (https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Turkish_WordList_10K), 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. However, previous studies have shown that type frequency yields better results in modelling phonological intuitions (Hayes and Wilson 2008:395). The current study leaves this alternative strategy for future investigation.

[4]The concept of O/E can be traced back to Fisher (1922)'s "On the mathematical foundations of theoretical statistics".

to discover phonotactic constraints (Pierrehumbert, 2001a; Frisch et al., 2004). However, this equation assumes an empty hypothesis grammar, which becomes inaccurate once any constraint is added, as discussed in Wilson and Obdeyn (2009) and Wilson (2022).

The current criterion O/E draws inspiration from Hayes and Wilson (2008), while a crucial ratio 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:

$$O[C] = |\{s \in S : C \in \texttt{factor}(s, 2)\}| \tag{3.2}$$

In a toy sample $S =$ {CVC, CVV, VVC, VVV, VCV, CCV}, O[*CC] $= 1$, O[*CV] $= 4$, O[*VC] $= 3$, O[*VV] $= 3$. Here, O[*VV] is 3 rather than 4 because, by definition, $O[C]$ 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[*VV] 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 $E[C]$ represents the number of unique strings in the hypothesized language L that violate C, under a noncumulative hypothesis grammar G.[5] Following Hayes and Wilson (2008), the current study works with an estimation to $E[C]$ by limiting the maximum string length in L to ℓ_{max}, mirroring the length of the longest string in the input data S. $E[C]$ is then approximated by:

$$E[C] \approx \sum_{\ell=1}^{\ell_{\text{max}}} E_\ell[C] \tag{3.3}$$

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

[5]Hayes and Wilson (2008:427) provides a method to estimate E for cumulative constraints, where a string can violate one constraint multiple times.

$$E_\ell\left[C\right] = |S_\ell| \times \textit{Ratio}(C, G, \ell) \tag{3.4}$$

$\textit{Ratio}(C, G, \ell)$ represents the proportion of strings of ℓ length accepted by G but violating C. This is found by comparing the accepted strings in G and $G' = G \cup \{C\}$, where C is added to G.[6]

$$\textit{Ratio}(C, G, \ell) = \frac{\textit{Count}(G, \ell) - \textit{Count}(G', \ell)}{\textit{Count}(G, \ell)} \tag{3.5}$$

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

Table 3.1 illustrates this calculation with exception-free input data that perfectly align with each hypothesis grammar G. The first row shows an empty hypothesis grammar ($G = \emptyset$) along with input data {CCC, CCV, CVC, CVV, VVV, VCV, VCC, VVC} (where $|S_3| = 8$). $\textit{Count}(\emptyset, 3) = 8$, given that the empty hypothesis grammar permits eight potential strings {CCC, CCV, VCC, CVC, CVV, VVV, VCV, VVC} of length 3.

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

Table 3.1: The list of idealized input data and corresponding hypothesis grammar, as well as expected frequencies for length 3; the input data S_3 here is idealized and identical to the target language L_3.

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

[6]Efficient computation can be done using a short-distance algorithm on finite-state automata, such as `shortestdistance` in pynini (Gorman, 2016). The author acknowledges Colin Wilson for guidance on the implementation of this algorithm.

calculated as follows:

$$
\begin{aligned}
E\,[\text{*CC}] &= E_3\,[\text{*CC}] \\
&= |S_3| \times \textit{Ratio}(\text{*CC}, \emptyset, 3) \\
&= 8 \times (\frac{\textit{Count}(\emptyset, 3) - \textit{Count}(\{\text{*CC}\}, 3)}{\textit{Count}(\emptyset, 3)}) \\
&= 8 \times (\frac{8-5}{8}) \\
&= 3
\end{aligned}
\tag{3.6}
$$

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

Following this update, ungrammatical strings (violating G) are filtered from the input data S. When G becomes {*CC}, as shown in the second row of Table 3.1, the input data shrinks to {CVC, CVV, VVV, VCV, VVC} ($|S_3| = 5$). $E[\text{*CC}]$ drops to zero, because *CC is already penalized by G (*CC $\in G$). In other potential constraints, for example, $E[\text{*VV}] = |S_3| \cdot (\frac{5-2}{5}) = 5 \cdot \frac{3}{5} = 3$, three of the five strings allowed by $G = \{\text{*CC}\}$ violate *VV.

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 θ_{max}, also known as the *accuracy schedule* (Hayes and 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 $\theta_{\text{max}} = 1$. This structure prioritizes the integration of potential constraints with the lowest O/E values.[7] θ_{max} can be interpreted as follows: the higher

[7]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 and Hayes, 2002, 2003b; Hayes and Wilson, 2008). α is set to 0.975 after Hayes and 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 prioritize potential constraints where the O and E disparity is high.

θ_{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 θ_{max}.

Equipping the Exception-Filtering learner with the accuracy schedule adapted from Hayes and 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 θ_{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 dissertation, θ_{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. The iterations of hyperparameter tuning is determined by the intervals, e.g. ten iterations if the accuracy schedule is [0.1, 0.2, 0.3, ..., 1].

Future psycholinguistic studies are required to better model the factors that determine θ_{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 θ_{max} in the Exception-Filtering learner.

3.5 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 formalized as follows: given the input data S, select a hypothesis grammar G from the hypothesis space, so that G approximates the target grammar $\mathcal{T}$ that defines the

target language $\mathcal{L}$.[8] 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 $\mathcal{T} = \{$*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 visualized in Figure Figure 3.3, given the input data S, tier, and the maximum O/E threshold θ_{max}, the learner first initializes 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.[9]

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

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

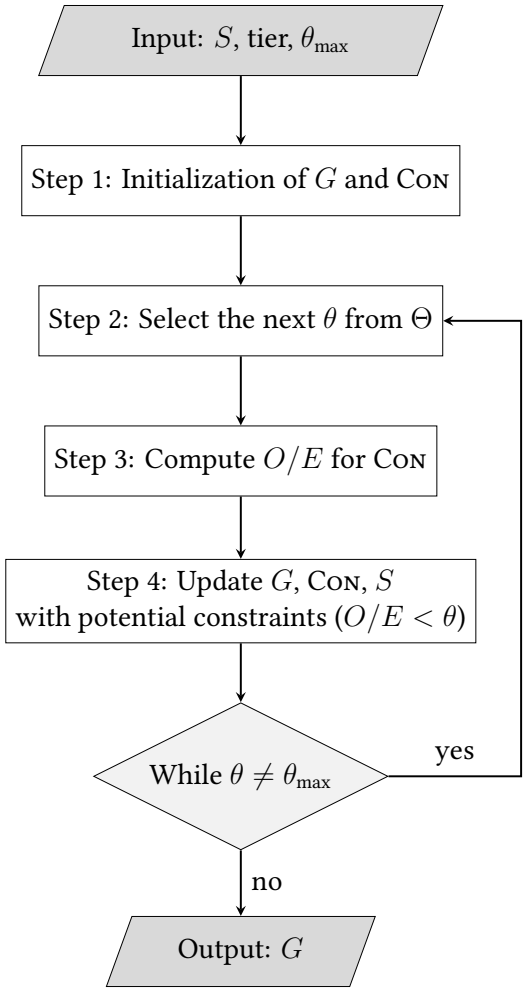


Figure 3.3: The learning procedure of the Exception-Filtering learner.

3.5.1 Step 1: Initialization

Given the input data S and tier {C, V}, the learning process begins with the initialization 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 initialization process is shown in Table 3.2, where the left side shows the initialization 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 3.2: Initialization.

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

Following the initialization, 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 $G' = \{$*CC$\}$. The ratio that a string violates *CC in the sample is *Ratio*$($*CC$, \emptyset, 3) = 1 - \frac{5}{8} = \frac{3}{8}$. As a result, $E[$*CC$] = |S| \cdot$ *Ratio*$($*CC$, \emptyset, 3) = 6 \cdot \frac{3}{8} = 2.25$, as illustrated in Table 3.3.

	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 3.3: Compute O and E.

3.5.3 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 3.4. 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 3.4.

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

Table 3.4: 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.[10] 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.

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

3.5.4 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 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 3.5), 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 learned grammar matches the target grammar $\mathcal{T} =$ {*CC}, as shown in Table 3.5.

.

	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 3.5: Step 2 and 3 after the first iteration.

3.6 Summary

To summarize, 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

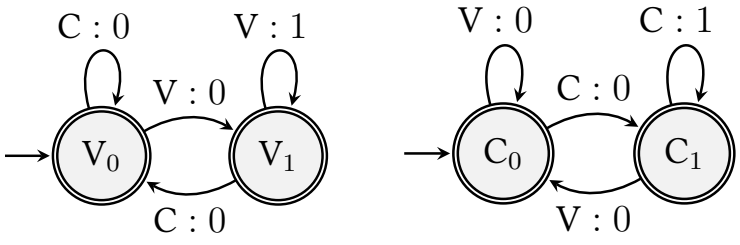
and expected type frequency. The learner iteratively filters out lexical exceptions from the input data, rather than accepting them in the hypothesis grammar.

3.7 Formal algorithms

This section describes the technical details regarding the calculating of O and E in the Exception-Filteirng Phonotactic Learner.

3.7.1 Weighted Finite-state Automata

k-factors, as a subregular language (Rogers and Pullum, 2011), can be characterized by deterministic Weighted Finite-state Acceptors (WFAs), which is essential for the computation of expected frequency in the current proposal. Weighted finite-state acceptors can represent possible strings in a formal language and encode phonotactic grammars. For the convenience of discussion, I assign $w = 0$ for *allowed* transitions and $w = 1$ for *penalized* transitions. WFAs of $\{C_1 : {}^*\text{VV}\}$ and $\{C_2 : {}^*\text{CC}\}$ in Figure 3.4. For example, in $\mathcal{M}_1$, the transition from state V_1 after accepting V is penalized ($w = 1$), which encodes the constraint *VV. In the intersected WFA $\mathcal{M}$, after accepting V, the transition of V from state V also receives penalty $w = 1$.



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

Figure 3.4: $\mathcal{M}_1$ for $\{C_1 : {}^*\text{VV}\}$ and $\mathcal{M}_2$ for $\{C_2 : {}^*\text{CC}\}$

The *composition* of WFAs (for a detailed definition, see Mohri et al., 2002, P.6) can be used to update hypothesis grammar. $\mathcal{M} = \mathcal{M}_1 \circ \mathcal{M}_2 \circ \ldots \circ \mathcal{M}_n$ is the composition of WFAs that embodies the combination of individual constraints in the hypothesis grammar $G = \{C_1, \ldots, C_n\}$. In the

learning procedure, every new constraint C corresponds to a WFA $\mathcal{M}_C$, and $\mathcal{M}' = \mathcal{M} \circ \mathcal{M}_C$ is the new WFA for the updated hypothesis grammar $G' = G \cup \{C\}$. Consider $\mathcal{M}_1$ the WFA of the original grammar {*VV}, and $\mathcal{M}_2$ encodes a newly added constraint {*CC}. Figure 3.5 shows the composition of $\mathcal{M}_1$ and $\mathcal{M}_2$ that correspond to the grammar {*CC, *VV}. In this dissertation, the transition weights on the WFAs represent negative log probabilities.

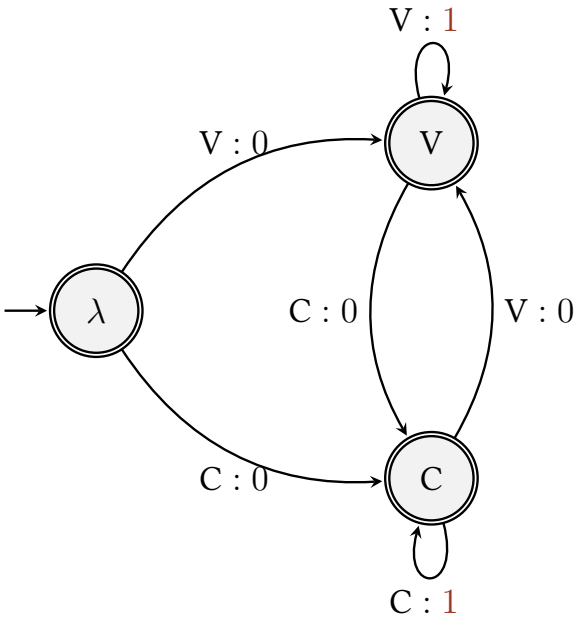


Figure 3.5: The minimized intersected $\mathcal{M}_1$ and $\mathcal{M}_2$

A braid $\mathcal{B}_\ell$ is a WFA that accepts Σ^ℓ. As shown in Figure 3.6, when $\ell = 3$, only the states following the paths of exactly three transitions C_3 and V_3 are accepting states in $\mathcal{B}_3$. the state indices indicate the symbol from the last transition and the number of previous transitions. For example, V_3 can only be reached after three transitions from the starting state λ, and immediately after the machine reads the symbol V.

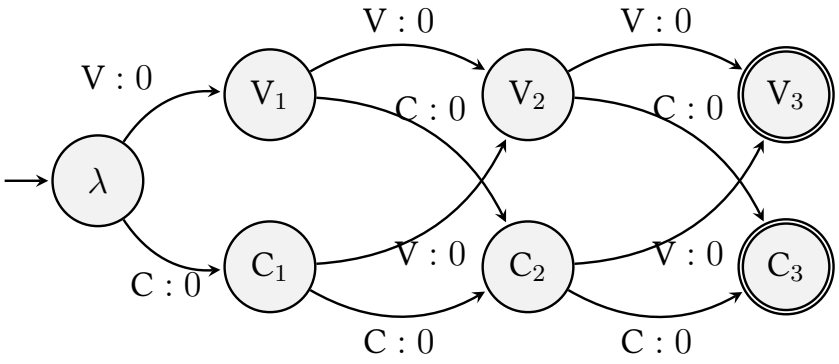


Figure 3.6: $\mathcal{B}_3$ that accepts all possible 3 length strings

The composition of WFAs can also be used to represent possible strings of certain lengths for a given grammar. As illustrated in Figure 3.7, $\mathcal{N}_\ell = \mathcal{B}_\ell \circ \mathcal{M}$ is the composition of $\mathcal{B}_\ell$ and the WFA $\mathcal{M}$, which accepts all possible strings s of length ℓ. In $\mathcal{N}_3$, the paths of allowed strings receive $w = 0$ on every transition, such as $\lambda \xrightarrow{\text{V:0}} \text{V}_1 \xrightarrow{\text{C:0}} \text{C}_2 \xrightarrow{\text{V:0}} \text{V}_3$ (VCV), while the path of a penalized string receives at least one $w = 1$, such as $\lambda \xrightarrow{\text{V:0}} \text{V}_1 \xrightarrow{\text{C:0}} \text{C}_2 \xrightarrow{\text{C:1}} \text{C}_3$. This means the path of VCC is penalized by the current hypothesis grammar encoded in the WFA $\mathcal{N}_3$.

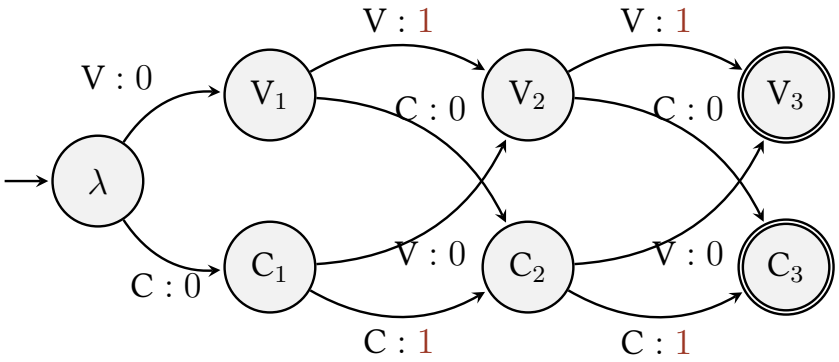


Figure 3.7: $\mathcal{N}_3 = \mathcal{B}_3 \circ \mathcal{M}$

3.7.2 Shortest-distance algorithm

To find the grammatical strings in a grammar G, we use the shortest distance algorithm on the WFA. This algorithm finds the minimum-weight path from the initial state to all other states in

the automaton. Since the weights of paths corresponding to grammatical strings are set to zero, the shortest distance algorithm effectively returns the paths that represent the sequences that satisfy G.

1. Initialize a distance function $d : Q \to \mathbb{R} \cup \{+\infty\}$ that maps each state to a numerical value. Set $d(q_0) = 0$ for the initial state q_0, and $d(q) = +\infty$ for all other states.

2. For each state q in Q, and for each arc (q, q', w) in A where q' is a destination state from state q with weight w, if $d(q') > d(q) + w$, then set $d(q') = d(q) + w$.

3. Repeat step 2 for $|Q| - 1$ times, where $|Q|$ is the number of states in the FSA.

Algorithm 1: Shortest Distance in Weighted FSA Algorithm

Input : Set of states Q, set of arcs $A = \{(q, q', w) | q, q' \in Q, w \in \mathbb{R}^+\}$

Output Distance function $d : Q \to \mathbb{R} \cup \{+\infty\}$

:

Initialize d such that $\forall q \in Q, d(q) = +\infty$ and $d(q_0) = 0$ for the initial state q_0;

for $|Q| - 1$ *iterations* **do**

  forall $q \in Q$ **do**

    forall $(q, q', w) \in A$ **do**

      if $d(q') > d(q) + w$ **then**

        Update $d(q')$ to $d(q) + w$;

      else

        Continue;

      end

    end

  end

end

The algorithm for computing O is formalized as follows:

Algorithm 2: COMPUTE_O

input : The set of constraints $\textsc{Con} = \{C_1, \ldots, C_n\}$ and the learning sample S

output: A dictionary of constraint-value mappings

initialization: $O \leftarrow \{C_1 : 0, \ldots, C_n : 0\}$;

for $C \in \textsc{Con}$ **do**

  for $s \in S$ **do**

    if $C \in \texttt{factor}(s, 2)$ **then**

      $O[C] \leftarrow O[C] + 1$;

The algorithm for approximating E is formalized as follows:

Algorithm 3: APPROXIMATE_E

input : The set of constraints $\textsc{Con} = \{C_1, \ldots, C_n\}$, the hypothesis grammar G, and the learning sample S

output: A dictionary of constraint-value mappings

initialization: $E \leftarrow \{C_1 : 0, \ldots, C_n : 0\}$, WFA $\mathcal{M}$ for G;

for $\ell \in \{1, \ldots, \ell_{\text{max}}\}$ **do**

  initialize $\mathcal{B}_\ell$;

  $\mathcal{N}_\ell \leftarrow \textsc{compose}(\mathcal{B}_\ell, \mathcal{M})$;

  for $C \in \textsc{Con}$ **do**

    if $C \notin G'$ **then**

      $\mathcal{M}' \leftarrow \textsc{compose}(\mathcal{M}, \mathcal{M}_C)$;

      $\mathcal{N}'_\ell \leftarrow \textsc{compose}(\mathcal{B}_\ell, \mathcal{M}')$;

      $E[C] \leftarrow |S_\ell| \times (1 - (Z(\mathcal{N}'_\ell)/Z(\mathcal{N}_\ell)))$

    else

      $E[C] \leftarrow 0$

CHAPTER 4

EVALUATION OF THE EXCEPTION-FILTERING PHONOTACTIC LEARNER

This chapter evaluates the core proposal of the Exception-filtering Phonotactic Learner in realistic corpora.

4.1 Correlation Tests

This section illustrates 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 learned grammar?
3. Model assessment: Evaluating the performance of the model with new data. This is achieved through the statistical tests against test dataset as discussed below;
4. Model comparison: Comparing the performance of different models.

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

Moreover, the current proposal is compared to the learning algorithm proposed by Hayes and Wilson (2008, henceforth HW learner) due to its widespread acceptance in the field and its accessible software (UCLA Phonotactic Learner; https://linguistics.ucla.edu/people/hayes/Phonotactics/), making it an ideal benchmark for comparison. In the case studies, the hyperparameters Max O/E (0.1 to 1) and Max gram size (2 to 3) in the HW learner were fine-tuned so that only the highest performing models across all tests are reported.[1] A 300 Maximum constraint limit was only established in the Turkish case study due to hardware limitations when handling a large corpus. Moreover, the default Gaussian prior is used to reduce overfitting and handle exceptions Hayes and Wilson (2008:387; $\mu = 0, \sigma = 1$; see more discussion on this exception-handling mechanism in §4.5).[2]

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

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

Although none of the learning models here claim to be the exact algorithm performed by child learners, comparing their learning results and behavioural data provides valuable insights into the underlying mechanisms of phonotactic learning in the face of exceptions. In English and Polish case studies, the learned grammars are tested on the acceptability judgments from

[1] Similarly, $\theta_{\max}$ in the Exception-Filtering learner is also reported on the best-performance basis.

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

behavioural data. In the Turkish case study, while conducting a new experiment falls outside the scope of the current study, the study approximates the acceptability judgments utilising the experimental data from Zimmer (1969). This is in line with the methodology employed by Hayes and Wilson (2008) for deriving acceptability judgments in English from Scholes (1966). Moreover, the learned grammar is contrasted with the documented grammar as analysed by human linguists. This has been a standard method in phonotactic modelling. For instance, Hayes and Wilson (2008) compared the learned grammars of Shona and Wargamay with the phonological generalizations in the previous literature. Gouskova and Gallagher (2020) used a method to generate grammaticality labels for nonce words based on phonological generalizations that are experimentally verified (§4.4).

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

4.1.1 Correlation Tests

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

Pearson's r requires the assumption of linearity, positing that intervals between ratings are of equal size (e.g., the distance between 1 and 2 is the same as between 4 and 5). However, this assumption may not hold for Likert ratings (Gorman, 2013; Dillon and Wagers, 2021), even if they are averaged over participants. Moreover, the Pearson correlation test also requires both variables to be continuous and their relationship to be normally distributed. The categorical grammaticality predicted in the current proposal does not satisfy this requirement. Therefore, Pearson's r is not reported in this study.[3]

Non-parametric tests measuring rank correlations are more appropriate as they make weaker

[3]Consequently, the *temperature* parameter in Hayes and Wilson (2008:400) is omitted, which only plays a role in their Pearson's correlation test and linear regression.

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

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

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

4.1.2 Classification Accuracy

When categorical grammaticality labels are provided in the test data, this paper utilizes *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 learned MaxEnt grammar into categorical grammaticality judgments (Hayes and Wilson, 2008: 385). Specifically, sequences with harmony scores equal to or below a certain threshold were classified as grammatical, whereas those with harmony scores exceeding the threshold were classified as ungrammatical. The optimal threshold was chosen, from the minimum to the maximum of all harmony scores, to maximize the binary accuracy of the learned MaxEnt grammar. In other words, the current proposal is compared to the maximal performance that a MaxEnt grammar can achieve in binary accuracy.

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

4.2 Case Study: English Onsets

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

4.2.1 English Input Data

The input of the learner is a "modestly" exceptionful input data, which consists of word-initial clusters taken from 31,985 distinctive word types drawn from the CMU Pronouncing Dictionary. Each of these words has been encountered at least once in the CELEX English database (Baayen

et al., 1995; Hayes and Wilson, 2008; Hayes, 2012). This methodology is designed to mirror the learning experiences of children (Pierrehumbert, 2001b).

There are 90 unique onsets in the input data. Table 4.1 illustrates how the majority of the input data (31,641 to be precise) are classified as nonexotic (Table 4.1a), while the onsets of 344 words are considered exotic (Table 4.1b) per Hayes and Wilson (2008). The HW learner yields worse performance when exposed to input data with "exotic" items compared to samples containing only nonexotic items. The current study claims that some, if not all, of these exotic items are lexical exceptions, especially those sequences borrowed from other languages, such as [zl] *zloty* from Polish. Following Hayes and Wilson (2008:395), [Cj] onsets are removed from the corpus due to considerable phonological evidence indicating that the [j] portion of [Cj] onsets is better parsed as part of the nucleus and rhyme, e.g., *spew* is analysed as [[sp]onset [ju] rhyme][4]. This filtering of [Cj] onsets leads to the input data characterized as "modestly exceptionful" because there are only few remaining exotic onsets.

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

k	2,764	w	780	s p	313	θ	173	ʃ r	40	f j	55	ʃ m	5	z j	2
r	2,752	n	716	f l	290	s w	153	s p l	27	m j	54	n j	4	h r	1
d	2,526	v	615	k l	285	g l	131	ð	19	h j	50	s k j	4	m w	1
s	2,215	g	537	s k	278	h w	111	d w	17	k j	45	ʃ n	4	n w	1
m	1,965	d͡ʒ	524	j	268	s n	109	g w	11	p j	34	b w	3	p w	1
p	1,881	s t	521	f r	254	s k r	93	θ w	4	b j	21	ʃ t	3	s r	1
b	1,544	t r	515	p l	238	z	83	s k l	1	d j	9	ʃ w	3	s θ	1
l	1,225	k r	387	b l	213	s m	82			t j	6	ʒ	3	ʃ p	1
f	1,222	ʃ	379	s l	213	θ r	73			v j	6	f w	2	v r	1
h	1,153	g r	331	d r	211	s k w	69			s f	5	g j	2	z l	1
t	1,146	t͡ʃ	329	k w	201	t w	55			s p j	5	k n	2	z w	1
p r	1,046	b r	319	s t r	183	s p r	51			ʃ l	5	v l	2		

(a) Nonexotic input data

(b) Exotic input data

Table 4.1: Type frequency of English onsets in the input data

Several phonotactic patterns are worth noting while interpreting the learned grammar, especially whether the attested "exotic" onsets such as [sf, zl, zw] are deemed ungrammatical. Moreover, previous studies have emphasized the impact of the Sonority Sequencing Principle (SSP) on English phonotactic judgments. According to the SSP, onsets featuring large sonority rises, such as "stop + liquid" combinations (e.g., [pl, bl, dr]), are generally favoured as being well-formed (Daland et al., 2011).[5] The current study only uses the SSP to better interpret the learned grammar. Capturing the effects of the SSP on unattested clusters, also known as *sonority projection* (Daland et al., 2011; Jarosz and Rysling, 2017), would require featural representations, which is beyond the scope of this paper.

[5]This paper assumes the conventional sonority hierarchy: stops « affricates « fricatives « nasals « liquids « glides (Clements, 1990), and discusses alternative hierarchy from Rubach and Booij (1990) in the Polish case study (§4.3).

4.2.2 Learning Procedure and Learned Grammar

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

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

Table 4.2: A grammar learned from the English sample. The first symbol of a two-factor sequence is denoted by the left column, while the second symbols are represented by segments on the penultimate top row. Cells highlighted indicate the attested two-factors within the input data, while blue and red separately correspond to grammatical and ungrammatical two-factors.

The Exception-Filtering learner learns consistent categorical grammars in every simulation, owing to the discrete nature of constraint selection. Arranged according to the sonority hierarchy, Table 4.2 illustrates the learned grammar when the maximum threshold $\theta_{\max}$ is set at 0.1, which delivers the optimal performance during the evaluation. The left column denotes the first symbol in a two-factor, while the penultimate top row represents the second symbol. The learner deems

grammatical two-factors, such as [pl], as 1, and ungrammatical ones, such as [pt], as 0. The grammatical two-factors such as [bl] in the learned grammar are all attested, while the attested ungrammatical two-factors such as [pw] indicate detected lexical exceptions. The $\theta_{\text{max}} = 0.1$ 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 learned grammar yields several interesting insights. Only clusters with large sonority rises are permitted by the learned 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 learned grammar.

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

4.2.3 Model Evaluation in English

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

[ʃl] in *schlep* (Daland et al. 2011:203).

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

Table 4.3 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 learned 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 4.4 provides a performance comparison among the Exception-Filtering ($\theta_{\max} = 0.1$), Baseline, and HW learner (Max O/E = 0.3, Max gram = 2, the same as Hayes and Wilson, 2008). Correlation scores are compared across the entire test dataset as a whole. It should be noted that the test dataset from Daland et al. (2011) excludes several exceptional onsets penalized by the Exception-Filtering learner, such as [*sf].

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

Table 4.4: Results of the best performance in Exception-Filtering, Baseline, and HW learner; correlation tests are reported with respect to averaged likert ratings in English; best scores are underscored

The reported correlation scores of all models are significantly different from zero at a two-tailed alpha of 0.01. Both the Exception-Filtering and Baseline learners delivered comparable

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

Table 4.3: Type frequency, averaged Likert ratings, and predicted grammaticality by the learned grammar of English nonce word onsets; detected exceptions (nonzero frequency and $g = 0$) are highlighted; sorted by averaged Likert ratings

performances[6], while the HW learner demonstrated slightly superior results, especially in terms of Spearman's ρ and Kendall's τ. Interestingly, the close-to-one Goodman and Kruskal's γ observed in both Exception-Filtering and Baseline learners indicates a higher number of tied pairs in nonparametric tests, leading to a marginally reduced Kendall's τ.

Although the Exception-Filtering learner shows a comparable performance on par with other well-established models, it did not stand out in approximating the acceptability judgments

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

of Daland et al. (2011). However, the relatively modest performance of the Exception-Filtering learner in the modestly exceptionful input data sets the stage for improved learning results in the forthcoming sections dealing with highly exceptionful data.

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

4.3 Case Study: Polish Onsets

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

4.3.1 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.

	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			
Palatalized Velar	kʲ, gʲ					

Table 4.5: Polish consonant inventory (derived from the input data)

Table 4.5 shows the consonants that appear in the input data. The current study uses a uniform system for converting orthography to IPA, remaining neutral on the ongoing debate surrounding the specific phonetic properties of certain segments, particularly the retroflex consonants *cz* [t͡ʂ], *drz*/*dż* [d͡ʐ], *sz* [ʂ], and *rz*/*ż* [ʐ] (Jarosz and Rysling, 2017; Kostyszyn and Heinz, 2022). Polish is known for allowing complex onsets (up to four consonants such as [vzdw]) that defy SSP (Jarosz, 2017; Kostyszyn and Heinz, 2022)[7]. 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.

4.3.2 Learning Procedure and Learned 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 initializes possible constraints for 30 consonants that appear

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

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 learned grammar. After the learning process, Table 4.6, arranged according to the sonority hierarchy, illustrates the learned grammar when $\theta_{\max}$ is set at 0.1, which delivers the optimal performance.

The learned grammar provides intriguing information on the attested SSP-defying onsets(Jarosz, 2017). Most grammatical two-factors that violate the SSP are obstruent pairs such as "fricative + stop", "fricative + stop", and "fricative + fricative". Rubach and Booij (1990) proposed that stops, affricates, and fricatives have indistinguishable sonority and should be considered as a single category, "obstruents", in the context of the SSP. If one follows this proposition and disregards obstruent initial onsets, most of the remaining two SSP-defying factors, such as "nasal + obstruent" [r] and "glide + stop" [wd], have relatively low type frequencies and are deemed ungrammatical by the learned grammar. Only 4 of 900 two-factors are grammatical while defying SSP (have a low or equal sonority rise), namely [lv, rv, mn, mɲ]. In essence, while a comprehensive evaluation of SSP's role in phonotactic learning is beyond the scope of this study, it is noteworthy that the learned grammar here shows a viable approach to interpreting SSP-defying onsets in the context of lexical exceptions.

.

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

Table 4.6: Learned gammar from Polish input data. The first symbol of a two-factor sequence is denoted by the left column, while the second symbol is represented by segments on the penultimate top row. Cells highlighted indicate the attested two-factors within the input data, while blue and red separately correspond to grammatical and ungrammatical two-factors.

4.3.3 Model Evaluation in Polish Data

This section evaluates the degree to which the learned grammar reflects acceptability judgments gathered from experimental data in Polish. The test dataset consists of 159 nonce words, which are constructed from a combination of 53 word-initial onsets (heads) and 3 trisyllabic VCVC(C)V(C) tails. The test dataset also includes 240 attested fillers, varying in word length (1 to 4 syllables) and onset length (0 to 3 consonants). This setting allows for the evaluation of the learner's performance on both attested and unattested sound sequences. Likert ratings were collected from 81 L1 Polish-speaking adults through an online experiment conducted on Ibex Farm (Jarosz and Rysling, 2017).

Table 4.7 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 learned grammar. Exceptions detected by the learned grammar (nonzero frequency and $g = 0$) are highlighted.[8] For instance, [zj] is deemed ungrammatical, which is reflected in its average score of 2.259 on a 1 to 7 Likert scale.

Table 4.8 shows the correlation with respect to averaged Likert ratings in Table 4.7. The correlation scores are compared across the entire test dataset as a whole.[9] Correlations in all models significantly differ from zero at a two-tailed alpha of 0.01. In all correlation tests, the Exception-Filtering learner modestly outperforms the Baseline learner. It performs comparably to the benchmark HW learner, with a modestly lower Spearman's ρ and a modestly higher Kendall's τ.

[8]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.

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

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

Table 4.7: Type frequency, averaged Likert ratings, and predicted grammaticality by the learned grammar of Polish onsets; detected exceptions onsets are highlighted; sorted by Likert

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

Table 4.8: Results of the best performance in Exception-Filtering, Baseline, and HW learner; correlation tests are approximating averaged Likert ratings in Polish; categorized based on attestedness; best scores are underscored.

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

4.3.4 AIC and BIC

For a robust comparison of the learning models, Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are reported (Akaike, 1974; Schwarz, 1978; Anderson and Burnham, 2004; Burnham and Anderson, 2004; Shih, 2017; Keane et al., 2017; Wilson, 2022).[10] These criteria can be calculated for models with either binary or continuous predictors.[11] Both AIC and BIC aim to strike a balance between model fit and complexity, with lower values indicating models that provide better explanations of the data while minimizing complexity. In this study, model complexity is controlled by using the same regression model structure, only varying the predictors (i.e., the predicted scores from the Exception-filtering learner vs. the HW learner). This enables a fair comparison of the learning models under equivalent conditions.

Ordinal regression models are constructed to evaluate hypotheses about the role of predicted judgments (the predictor) from each learning model on the observed individual likert ratings (the outcome). Brant tests confirm that the proportional odds assumption holds for all regression models examined here. The result of ordinal regression models with respect to individual Likert ratings are reported in Table 4.9. The Exception-filtering learner achieves the highest regression slope and returns the lowest AIC and BIC values, which indicates that the categorical grammar learned a competitive fit to the data.

[10] There has been debate on the proper usage of AIC and BIC, see Burnham and Anderson (2004) for a detailed discussion.

[11] Given the likelihood L and the number of variables (k), $AIC = -2 * log(L) + 2 * k$, $BIC = -2 * log(L) + k * log(n)$. This is done using the `AIC()` and `BIC()` function in R.

		Exception-filtering	Baseline	HW
Ordinal regression	slope	1.45	1.27	1.42
	AIC	45,467.109	45,765.127	47,153.554
	BIC	45,519.353	45,817.370	47,205.798

Table 4.9: Ordinal regression with respect to individual likert ratings in Polish; best scores are underscored

To summarize, the Exception-Filtering learner, trained on Polish child-directed corpus, has illustrated its potential in extracting categorical grammars that approximate acceptability judgments. The performance of the model is on par with the HW learner in Spearman's ρ, and is modestly outperforms the benchmark HW learner and the Baseline learner in both Goodman-Kruskal's γ and Kendall's τ test, demonstrating its capability in approximating acceptability judgments. These results further substantiate the potential of the Exception-Filtering learner in inducing phonotactic patterns from realistic corpora.[12]

4.4 Case Study: Turkish Vowel Phonotactics

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

4.4.1 Turkish Vowel Phonotactics

This section applies the current proposal to vowel phonotactic patterns in Turkish. Turkish vowels are shown in Table 4.10. Turkish orthography is converted to IPA, including *ö* [ø], *ü* [y], and *ı* [ɯ].

[12]Additionally, individual subjects could be incorporated as a random effect in an ordinal mixed-effects regression model, also known as "cumulative link mixed model" (Hedeker and Gibbons, 2006; Christensen, 2019). This approach allows the model to accommodate for inter-individual variability in the ratings, thereby enhancing the precision of the model estimates. However, the implemented ordinal mixed-effects regression model `clmm(as.ordered(likert_rating) ~ score + (1|subj))` failed to converge during testing.

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

Table 4.10: Turkish vowel system

Turkish vowel phonotactic patterns are summarized 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 4.11 (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 /køj/ village" is combined with /lAr/, the resulting term is [køjler] "villages", demonstrating the nonlocal *ø…ɑ co-occurrence restriction. However, exceptions against this generalization exist both within roots and across root-affix boundaries, as illustrated in examples (c) and (d) in Table 4.11. 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)
	køj	køj-ler	“village”	
	jyz	jyz-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	‘grave”	(Inkelas et al., 2000)
	model	model-ler	“model”	
	silɑh	silɑh-lɑr	“weapon”	
	memur	memur-lɑr	“official”	
	sɑbun	sɑbun–lɑr	“soap”	
d.	etol	etol-ler	“fur stole”	(Göksel and Kerslake, 2004)
	sɑɑt	sɑɑt-ler	“hour, clock”	
	kɑhɑbɑt	kɑhɑbɑt-ler	“fault”	

Table 4.11: 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 4.12 (a). Table 4.12 provides examples of this pattern. Yet again, exceptions are noted, such as in the root [boğɑz] “throat” and its derived forms.[13]

[13]A unique case of exceptions is caused by the phenomenon of root-internal *labial attraction*, where ɑC_[+labial]u

	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”	
	køj	køj-y	køj-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”	
	jamuk	jamuğ-u	jamuğ-un	“trapezoid”	
	ymit	ymit-i	ymit-in	“hope”	

Table 4.12: Turkish round harmony patterns in morphophonological alternations (a) and exceptions (b) (Gorman, 2013)

Last but not the least, mid round vowels [ø] and [o] are typically restricted to initial positions in L1 Turkish words. This is evident in words like [ødev] “homework” and *ojun* “game”. Consequently, these vowels should not follow any other vowels, for example, *ɑ…ø and *e…o. However, in loanwords, mid round vowels may occur freely in any position.

Generally, a substantial number of exceptions to these phonotactic patterns arise from compounds and loanwords (Lewis, 2001; Göksel and Kerslake, 2004; Kabak, 2011). For example, the compound word [bugyn] “today” ([bu] “this” + [gyn] “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.

is produced given the intervocalic labial consonant, as seen in [sɑbur] “patient” (Lees, 1966). However, this pattern is not internalized by all L1 speakers, as shown in the ratings of nonce words by L1 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.

Despite many exceptions, these generalizations are not only well-documented in the literature, including Underhill (1976:25), Lewis (2001:16), Göksel and 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; Hohenberger 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 (Harrison and Kaun, 2000; 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 generalizations from both roots and derived forms, despite the high level of lexical exceptions within the input data?

4.4.2 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 $\approx$ 66000 roots and the elicited derived forms (root + affixes) produced by two adult L1 Turkish speakers.[14] Table 4.13 shows the type frequency of all nonlocal two-factors on the vowel tier in TELL. Two-factors that follow the Turkish vowel phonotactics introduced above are highlighted. This adult-direct corpus is

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

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

Table 4.13: 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 initializes possible constraints for 8 Turkish vowels, which yields 64 two-factors in the hypothesis space. The optimal maximum O/E threshold is 0.5. The learned grammar is illustrated in the first test dataset below.

4.4.3 Model Evaluation

This section evaluates the learning models in two separate test datasets below supported by the experimental data in Zimmer (1969).

Zimmer (1969)'s experiment

In a binary wordlikeness task, Zimmer (1969) asked L1 adult speakers to select which of two nonce words, for example, temez-temaz, was "more like Turkish". Experiment 1 had 23 participants, and Experiment 2 had 32, all of whom were L1 adult speakers of Turkish. Table Table 4.14 and Table 4.15 illustrate the effects of backness and roundness harmony on the wordlikeness experiment carried out by Zimmer (1969). The numbers represent how many participants selected the corresponding nonce word, while the responses indicating "no preference" were excluded.

Experiment 1				Experiment 2			
Harmonic		Disharmonic		Harmonic		Disharmonic	
temez	19	temɑz	3	pemez	30	pemɑz	2
teriz	23	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 4.14: Effects of backness harmony on Zimmer (1969)'s wordlikeness experiment, from Gorman (2013)

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

Table 4.15: Effects of roundness harmony on Zimmer (1969)'s wordlikeness experiment, from Gorman (2013)

The First Test Dataset (Categorical Labels)

Based on the data in Zimmer (1969), 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 generalizations.[15] 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* speaker. However, these categorical labels are supported by Zimmer (1969)'s behavioural experiment; the majority of participants preferred the harmonic to disharmonic roots in a yes/no rating task, which provides the evidence for the psychological reality of Turkish vowel phonotactic patterns encoded in the first test dataset. In other words, the first test dataset aims to evaluate how well the learned grammar mirrors the categorical phonotactic judgments of the *majority* of participants in Zimmer (1969)'s experiment. This follows the common practice in previous computational studies when acceptability judgments of nonce words in the test dataset are not accessible. For instance, Gouskova and Gallagher (2020) manually labelled the categorical grammaticality of nonce words in the test dataset based on documented phonotactic generalizations supported by behavioural experiments (Gallagher, 2014, 2015, 2016).

Table 4.16 summarizes the tests of classification accuracy on the first test dataset with categorical labels. The Baseline learner miscategorized 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.

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

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

Table 4.16: Performance comparison of Exception-Filtering, Baseline, and HW learner in the first test dataset (categorical labels); best scores are underscored

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

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

(a) Exception-Filtering

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

(b) HW

Figure 4.1: Compare the learned grammars of (a) Exception-Filtering learner and (b) HW learner

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

total binary accuracy of 0.969. Figure 4.1 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.

Compared to phonotactic generalizations in Turkish, the learned 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 learned 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.[16]

The Second Test Dataset (Approximated Acceptability Judgments)

The purpose of the second test dataset is to demonstrate that the learned categorical grammar can approximate the acceptability judgments in the behavioural data. The second testing data includes 36 nonce words in Zimmer (1969), and takes the proportion of "yes" responses averaged across participants to approximate the acceptability judgments of L1 speakers. The data show a gradient transition from harmonic, e.g., [temez] receives $19/23 \approx 0.826$ to disharmonic words e.g., [temɑz] $3/23 \approx 0.130$. This method is similar to Hayes and Wilson (2008)'s approach to create gradient acceptability judgments from the Scholes (1966) experiment, following previous studies (Pierrehumbert, 1994; Coleman and Pierrehumbert, 1997).

Table 4.17 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 corre-

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

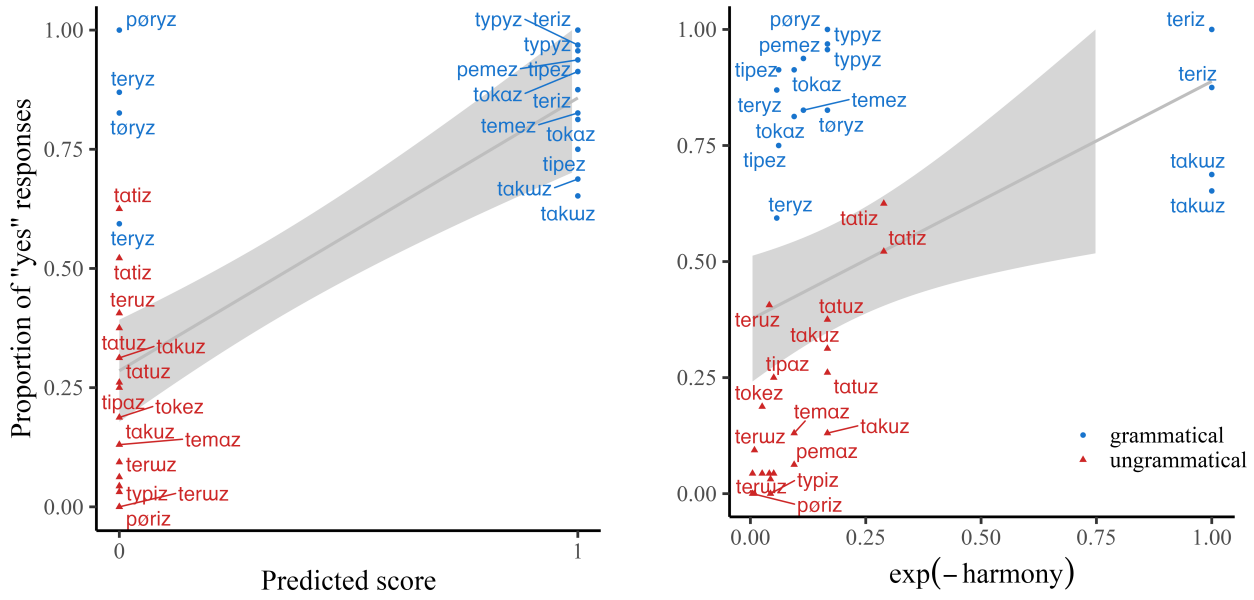
lations in all models differ significantly from zero at a two-tailed alpha of 0.01, the Exception-Filtering learner scored higher than the benchmark HW learner in all tests.

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

Table 4.17: Performance comparison of Exception-Filtering and HW learner in the second test dataset adapteds from Zimmer (1969)'s experiment; best scores are underscored.

Figure 4.2 visualizes the distribution of predicted score against the approximated acceptability in both Exception-Filtering and HW learner. Some words have two response rates as they appeared in two separate experiments. 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 judgments. The predicted scores of the Exception-Filtering learner cluster at 0 and 1, while the $\exp(-\text{harmony})$ is on a continuum.[17]

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



(a) Exception-Filtering

(b) HW

Figure 4.2: Scatterplot based on the learning results of two learners; expected grammaticality is hightlighted based on the documented phonotactic generalizations; some words have two response rates as they appeared in two separate experiments; overlapped words are omitted on the plots.

Both regression models reject the null hypothesis that the predicted judgments have no effect on the proportion of "yes" responses (Exception-Filtering: residual deviance $= 2.264, p < 0.001$; HW: residual deviance $= 4.073, p = 0.013$), at an alpha level of 0.05.

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

4.5 Summary

To summarize the case studies, in terms of interpretability and scalability, the categorical grammars learned in the case studies of English and Polish onset phonotactics largely align with the Sonority Sequencing Principle that penalizes most sequences with low sonority rises. The proposed learner also successfully generalized 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 learned grammars closely approximate the acceptability judgments observed in behavioural experiments and demonstrate competitive performance in model comparisons, highlighting the effectiveness of the exception-filtering mechanism.

This proposal 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 learned grammars with extragrammatical factors to model behavioural data, and marks initial steps in reassessing the ability of categorical grammars in approximating human judgments.

CHAPTER 5

DISCUSSION OF THE EXCEPTION-FILTERING PHONOTACTIC LEARNER

To summarize the case studies, in terms of interpretability and scalability, the categorical grammars learned in the case studies of English and Polish onset phonotactics largely align with the Sonority Sequencing Principle that penalizes most sequences with low sonority rises. The proposed learner also successfully generalized 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 learned grammars closely approximate the acceptability judgments observed in behavioral 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.

5.1 Extragrammatical Factors

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

In the current study, in tandem with the learned grammar, extragrammatical factors con-

tribute to acceptability judgments in behavioural experiments. For example, previous studies have shown that lexical similarity and frequency are significant predictors of acceptability judgments (Bailey and Hahn, 2001, 2005; Frisch et al., 2000). Performance-related variables, such as individual differences and task effects, can also influence acceptability judgments. Therefore, a comprehensive evaluation of a learned grammar against acceptability judgments 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.

.

5.2 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 misgeneralization of accidental gaps by adjusting the maximum threshold. However, this is not a fundamental solution and places an excessive burden on a simple statistical criterion.

A more principled solution to the challenge of accidental gaps is to incorporate feature-based constraints, as suggested by Wilson and Gallagher (2018). Segmental representations may overlook subsegmental generalizations—underrepresented segmental two-factors in the input data can exhibit high frequency in feature-based generalizations. 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 penalize [−approximant] after b, but allow b[+approximant], hence avoiding overpenalising accidental gaps with [bw] onsets. By considering the entire natural class, the grammar can recognize subsegmental patterns that are overlooked in segmental representation. As Hayes and 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 and Wilson (2008) and the bottom-up strategies proposed by Rawski (2021). The current study offers a straightforward demonstration of the concept here: consider a simplified feature system illustrated in Table Table 5.1, a feature-based Exception-Filtering learner initializes the most general feature-based potential constraints, e.g., *$[+F][+F]$, *$[+F][+G]$, etc.

	F	G
C	$+$	$+$
V	$+$	$-$

Table 5.1: Simplified feature system.

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

5.3 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 and Johnson, 2003; Berger et al., 1996) from input data. The HW learner adjusts constraint weights to maximize 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 underpenalization of lexical exceptions can compromise generalizations for nonexceptional candidates (Moore-Cantwell and Pater, 2016). For example, in the Turkish case study, the HW learner underpenalized the highly frequent disharmonic patterns such as ɑ...i in [tɑtiz] (Figure Figure 4.2). 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 and Wilson (2008).

This issue has motivated several interesting proposals to handle exceptions within the HW learner. Hayes and Wilson (2008:386) added a Gaussian prior to prevent overfitting by adjusting the standard deviation σ of the Gaussian distribution for constraint weights. 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 and Pater, 2016; Hughto et al., 2019; O'Hara, 2020). Lexically specific constraints such as *sf$_i$ would normally penalize the sequence [sf], except when it is in the indexed lexical exception *sphere*$_i$. In this way, the learned

grammar is able to allow exceptions without compromising the generalizations 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 regularization 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.

5.4 *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. 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 cri-

teria, 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 θ_{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 and 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 generalized 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.

5.5 Other Future Directions

The current study represents an initial step towards understanding the interplay between lexical exceptions and phonotactic learning. The primary objective of this study has been to address the issue of exceptions, rather than developing an all-encompassing learning model. This has led to significant simplifications in the proposed learning model. Therefore, the next step is to enhance the current proposal towards a more comprehensive model. First, this study uses a simplified noncumulative categorical grammar, while experimental evidence has indicated a cumulative effect on phonotactic learning (Breiss, 2020; Kawahara and Breiss, 2021). A 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 learned 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 and Heinz, 2016; Gouskova and Gallagher, 2020). Another promising direction is to extend the current approach to the hypothesis space defined by other formal languages (Jäger and Rogers, 2012).

CHAPTER 6

TOWARDS A TWO-STAGE PHONOTACTIC-ALTERNATION LEARNING MODEL

6.1 Introduction

As discussed in chapter 2, the phonological knowledge of human language learners encompasses PHONOTACTICS, the *syntagmatic* generalizations about the distribution of sound sequences in a language, and ALTERNATIONS, which refer to *paradigmatic* generalizations where the phonological form of a morpheme varies based on the (morpho-)phonological context. For instance, Turkish vowel backness harmony can be described either as a phonotactic constraint prohibiting front and back vowel co-occurrence (*[αback][$-\alpha$back]) or as an alternation that aligns the [back] feature of a vowel to that of its preceding vowel ([+syllabic]$\rightarrow$[αback]/[αback]_). These isomorphic generalizations yield different forms in plural suffixes as seen in [ip-**ler**] "rope-PL" versus [pul-**lar**] "stamp-PL".

However, phonotactics and alternation knowledge are not always isomorphic. On the one hand, some phonotactic constraints are not related to any observed alternations, such as the positional restriction in Turkish where no mid round vowels [o, ø] can follow another vowel, as described in chapter 8. On the other hand, some alternations are unrelated to any phonotactic restrictions, notably exhibited in the DERIVED-ENVIRONMENT EFFECTS (DEEs; Łubowicz, 2002; Burzio, 2011), where phonological alternations can exclusively occur, or be blocked, in morphologically-derived environment (Iverson and Wheeler, 1988; Kiparsky, 1993; Chong, 2019). For example, Chong (2019) shows that, while intervocalic [k] is often deleted across morpheme boundaries in Turkish, e.g., /bebek–In/ $\rightarrow$ [bebein], VkV sequence is statistically robust and the learning simulation based on Hayes and Wilson (2008) phonotactic learner cannot learn *VkV constraint from the naturalistic corpus Turkish Electronic Living Lexicon (Inkelas et al., 2000). In

sum, although alternation knowledge is constrained by phonotactics, they are distinct and should not be conflated in the phonological grammar. Therefore, although phonotactic constraints influence alternation patterns, they are inherently distinct phenomena within phonological grammar and should be treated as such in both processing and learning models.

6.1.1 The link between phonotactic and alternation learning

There has been converging evidence on the distinction and link between phonotactic and alternation learning (Pater and Tessier, 2006; Jarosz, 2011; Chong, 2021; Jo, 2024; Kuo, 2024; Jun et al., 2024). In particular, Chong's artificial grammar learning (AGL) experiments show that learners fail to learn alternations when they are not supported by stem phonotactics (experiment 1), while they also cannot readily generalize a learned static generalization to an unseen novel morphophonological alternation (experiments 2 and 3). The result of experiment 1 rejects models that ignore the role of phonotactics in phonological learning (Hale and Reiss, 2008; Reiss, 2017; Belth, 2023b), or learn phonotactic and alternation knowledge in completely independent mechanisms (Whang and Adriaans, 2017).[1] Meanwhile, the result of experiment 2 & 3 rejects models that learn phonotactic and alternation knowledge in a MONOLITHIC (as per Hayes, 2016) framework as in classic Optimality Theory (Prince and Smolensky, 1993; Smolensky, 1996).

Do and Yeung (2021) replicated Chong (2021)'s experiments with native speakers of Hong Kong Cantonese, but couldn't find the same facilitating effect of phonotactic learning in alternation learning. They concluded that the link between phonotactic-alternation learning was not found and previous positive results (Pater and Tessier, 2006; Pizzo, 2015; Pizzo and Pater, 2016; Chong, 2021) based on English speakers may be language-specific L1 effects. While more experimental replications on non-English speakers should be conducted in future studies, it is too early to reject the hypothesis on the link between phonotactics and alternation learning. For

[1] Whang and Adriaans (2017) proposed a two-stage model of phonotactics and alternation learning based on OT. However, in this proposal, the learned phonotactics are only combined with alternations in predicting surface forms after the alternation learning is complete.

example, Gong (2022)'s AGL experiment with Mandarin speakers couldn't replicate Do and Yeung (2021)'s null results. Moreover, in Do and Yeung (2021), the assumption that Hong Kong Cantonese speakers have no exposure to English is highly idealized, given its global influence. Therefore, language experience is not sufficient to account for the discrepancy between Do and Yeung (2021) and previous works. Moreover, it is possible that participants in Do and Yeung (2021) didn't learn phonotactics well enough to facilitate alternation learning. Under-training is a common issue in AGL settings because participants are exposed to merely a small amount of training data in a short period of time, compared to real-world language acquisition. Compare the blick test for testing phonotactics in Chong (2021, Figure 1a) and Do and Yeung (2021, Figure 2): in Do and Yeung (2021), the averaged rate of harmony answer is below 75% in harmony language, <50% in exceptional harmonic language (mixed), while >75% in harmony language, and >50% in exceptional harmonic (semi-harmonic) language in Chong (2017). Do and Yeung (2021, Table 3) also showed that no significant preference towards harmonic or disharmonic patterns in mixed language. However, this exceptional harmonic pattern represents a common scenario in natural languages such as Turkish vowel harmony, and a successful phonotactic learning in real-world learning should show preference to harmonic patterns. To summarize, there has not been a conclusive counter-evidence that falsifies the link between phonotactics and alternation learning shown in previous works (Pater and Tessier, 2006; Pizzo, 2015; Pizzo and Pater, 2016; Chong, 2017, 2021).

Given the converging empirical evidence, an important question arises from a learning perspective: How are alternations acquired such that they accurately predict surface forms following distinct phonotactic knowledge? An empirically grounded model should model phonotactic and alternation learning as *distinct* yet *interconnected* components, which is the methodology adopted in the current study.

While several previous proposals modeled phonotactic and alternating learning as distinct components, the connection between phonotactic and alternation learning is seldom explicitly

modeled. The most promising proposal was Albright and Hayes (2002), which incorporated a list of categorically illicit phonotactics (similar to k-factors introduced in chapter 3) into their rule-based learner. Albright and Hayes (2002)'s learner increases the statistical support for a morphophonological rule if it avoids phonotactically illicit surface forms. However, this model does not have a story of how these phonotactic constraints were learned from the training data.

6.1.2 The Structure of Phonotactic and Alternation Grammars

A phonotactic grammar can be mathematically characterized as is the set of illicit substrings of length k, also known as k-factors. This can be formalized by the function $\texttt{factor}(s, k)$, which generates all k-factors of a string s. For example, $\texttt{factor}(\text{iɑe}, 2) = \{\text{iɑ}, \text{ɑe}\}$, and $\texttt{factor}(\text{iɑɑ}, 2) = \{\text{iɑ}, \text{ɑɑ}\}$. A tier-based strictly k-local (TSL$_k$) grammar consists of all forbidden k-factors on a specific tier, known as TSL$_k$ constraints, such as {*ɑe, *ɑi, *iɑ, *øi, ...} over the vowel tier in Turkish. The *tier*, also referred to as a *projection* (Hayes and Wilson, 2008), is a subset of the inventory of phonological representations (e.g., consonants, vowels) that can be used to target the most relevant part of a string. This concept, although similar, is distinct from the traditional feature-based definition in Autosegmental Phonology (Goldsmith, 1976). In the context of nonlocal phonotactics, tiers often include only specific segments (henceforth "tier segments"), such as vowels. For example, as shown in Figure 6.1, a Turkish word [døviz] "currency" is represented as [øi] on the vowel tier. Nontier segments are ignored during the evaluation of tier-based constraints. A string is labelled as phonotactically grammatical if it does not contain any forbidden k-factors; otherwise, the string is considered phonotactically ungrammatical. Therefore, [døviz] is phonotactically ungrammatical in this example as it violates a tier-based 2-local constraint *øi on the vowel tier.

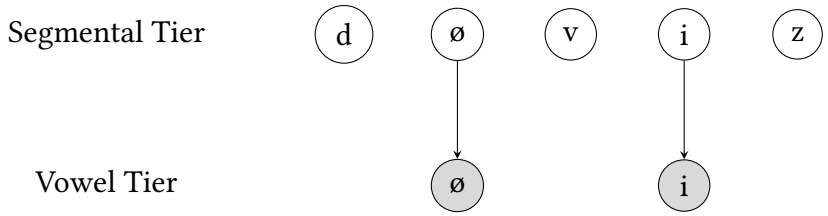


Figure 6.1: Extraction of vowel tier from the Turkish word [døviz] "currency". The vowel tier contains the vowels in this word, disregarding the non-tier consonants.

An alternation grammar can be mathematically characterized by a set of UR-to-SR mappings or *functions*, such as $\{$/iH/ $\rightarrow$ [ii], /uH/ $\rightarrow$ [uu], $\ldots\}$ or SPE-style rules $\{$H $\rightarrow$ i/i_, H $\rightarrow$ u/ u_, $\ldots\}$. Similar to TSL$_k$ languages, an Output Tier-based Strictly Local (OTSL$_k$) function keeps track of k-factors in the derived output of the preceding segment, assuming the derived output of the first segment is identical to its input UR. In other words, OTSL$_k$ functions rewrite URs of the most recent $k-1$ symbols, and predict the SRs based on these rewritten URs (output) in the rule application.

To elucidate the application of alternation grammars, consider an example where an OTSL$_2$ alternation grammar is applied to the UR /firmɑ-lAr-Hn/, representing "firm-PL-GEN", as shown in Table 6.1. The current study simplifies the alternation learning to target alternations in suffixes, and hence only apply the learned alternations to these domains during rule applications. The current study abstracts away from the *opacity* and *ordering* problem (for a comprehensive review see Baković, 2007). For every tier segment, the derivation exhaustively searches through the entire alternation grammar and does not progress to the next symbol unless all applicable rules have been applied.[2]

[2]During alternation learning, the learner assumes *maximum utilization* (Kiparsky, 1968), which favours patterns in which all rules are maximally utilized, and *transparency biases* (Kiparsky, 1971), which favours interactions in which processes are not opaque (Prickett, 2019), hence ignoring the rule ordering.

	/firmɑ-lAr-Hn/
A → [+back] / [+back] ___	firmɑ-l[-high, **+back**, -round]r-Hn
H → [-round] / [-round] ___	firmɑ-lɑr-[+high, 0back, **-round**]n
H → [+back] / [+back] ___	firmɑ-lɑr-[+high, **+back**, -round]n
	[firmɑ-lɑr-**ɯ**n]

Table 6.1: Derivation from UR to SR in a OTSL_2 grammar; changed features are highlighted

Given an alternation rule of the form $X \rightarrow Y / Z\_$, where X is the target symbol in UR, Y is the resultant symbol in SR, and Z represents the *preceding context* for the rule to apply, $X \rightarrow Y$ indicates the *structural change* from UR to SR, such as "H →i" (Albright and Hayes, 2003b). In this system, the *application conditions* of an alternation rule are defined as follows: For a target symbol B with the preceding context A in AB, B undergoes the structural change in $X \rightarrow Y / Z\_$ if and only if both (1) the target symbol B is entailed by X, and (2) the preceding context is entailed by Z. These conditions depends on the entailment relations between feature bundles, using the principle that a '0' value in an abstract entails all values ('+', '-', '0'), while specific values only entail themselves. Archiphonemes, like A and H, represents feature bundles and can be translated into sets of concrete segments. For instance, the archiphoneme H, characterized by the features [+high, 0back, 0round], encompasses the segments [i, u, y, ɯ]. In our example, H entails [i, u, y, ɯ, I, H], while I, with features [+high, -back, 0round], entails [i, y, I] but not [u, ɯ] due to their [+back] feature.

The application conditions based on entailment relations allow the general rules to be applied to more specific representations. Consider the derivation from /firmɑ-lAr-Hn/ to [firmɑ-lɑr-ɯn]. When the rule encounters the tier segment /A/: [-high, 0back, -round] within the -PL affix, the rule A → [+back] / [+back]___ is applied, transforming /A/ to [ɑ]: [-high, **+back**, -round]. As the alternation grammar moves to the next tier segment H: [+high, 0back, 0round] in /-Hn/, the

application of the rules H → [-round] / [-round] ___ and H → [+back] / [+back] ___ leads to a two-step transformation of H. Initially, H adopts the features [+high, 0back, -round], and then, considering the preceding context (ɑ, which was rewritten in the previous rule application), it changes to [+high, +back, -round].

The hypothesis spaces of phonotactic and alternation learning are shaped by the structure of corresponding potential hypothesis grammars. The hypothesis space for phonotactic learning includes various potential phonotactic constraint sets, whereas the hypothesis space for alternation learning encompasses a variety of input-output mapping sets. A key insight from computational learning theory emphasizes that an ideal hypothesis space should strike a balance: it needs to be sufficiently restrictive to facilitate efficient learning, yet broad enough to encompass the target grammar (Heinz and Riggle, 2011).

The current study employs TSL_2 as the hypothesis spaces for phonotactic learning ($\mathcal{H}_{\text{Phonotactics}}$). TSL_2 languages delineate a formally restrictive but typologically rich hypothesis space, capturing a broad spectrum of local and nonlocal phonotactics (Heinz et al., 2011) and alternations (Chandlee, 2014; Chandlee and Jardine, 2021). Experimental evidence from AGL experiments supports the viability of TSL_2 as a hypothesis space, indicating that adults can learn TSL_2 patterns, but struggled with patterns that fall outside the TSL_2 class (McMullin and Hansson, 2019). This is corroborated by formal language-theoretic research, which highlights the efficient learning properties of TSL_2 (Heinz et al., 2011; Jardine and Heinz, 2016; Jardine and McMullin, 2017). This approach has been successfully applied in previous work spanning both probabilistic and categorical approaches (Hayes and Wilson, 2008; Gouskova and Gallagher, 2020; Mayer, 2021; Dai et al., 2023; Heinz, 2007; Jardine and Heinz, 2016).

Similarly, $OTSL_2$ is employed as a hypothesis space for alternation learning ($\mathcal{H}_{\text{Alternation}}$), which captures *iterative spreading* of feature values in all vowel harmony patterns observed in Turkish, Finnish, and Hungarian, as explored in this study (Burness et al., 2021; Chandlee and Jardine, 2021).

In the current study, tiers are provided to the proposed learning model *a priori*, usually non-neutral vowels. Previous works have explored tier induction algorithms in phonotactic learning (Jardine and Heinz, 2016; Gouskova and Gallagher, 2020). Belth (2023b) proposes an innovative approach that discover tiers during phonological learning to encompass both local and nonlocal processes in alternation learning. A future direction is to incorporate these algorithms to the current Two-stage learning model.

6.1.3 Learning Problems

The learning problems of phonotactics and alternations are visualized in Figure 6.2 and Figure 6.3. The problem of learning phonotactics in the presence of lexical exceptions is formalized as follows: For input data S containing grammatical strings of the target language $\mathcal{L}$ and a finite set of ungrammatical strings not in $\mathcal{L}$ (lexical exceptions), the objective is to identify a grammar $G_{\text{Phonotactics}}$ within the hypothesis space $\mathcal{H}_{\text{Phonotactics}}$ such that $G_{\text{Phonotactics}}$ closely approximates (represented by $\rightsquigarrow$) the target grammar $\mathcal{T}_{\text{Phonotactics}}$ defining $\mathcal{L}$.[3]

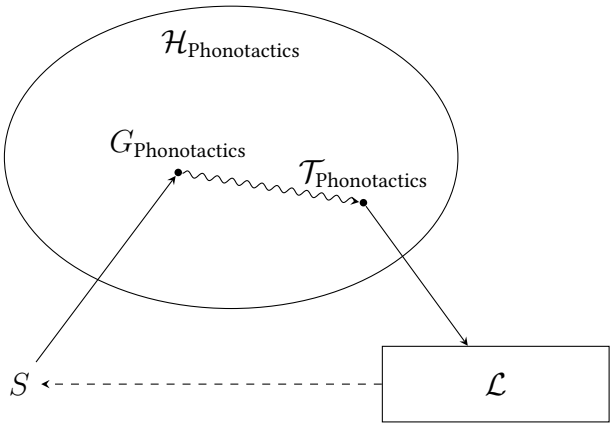


Figure 6.2: The phonotactic learning problems

[3]The assumption that a single uniform target grammar applies to all native speakers is a simplification that assumes no linguistic variations, where the input data should be generated by a single source, such as a parent-teacher. In a more realistic learning environment, there might be multiple target grammars across different speakers due to a variety of input data sources, causing variations among native speakers. However, the learning problem becomes learning a hypothesis grammar that approximates a target language defined by the intersection of target grammars.

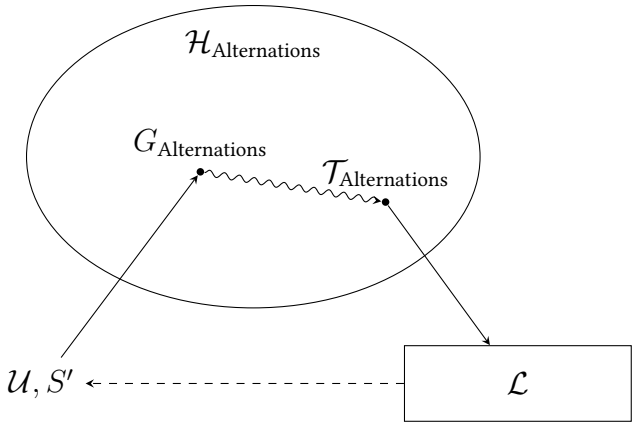


Figure 6.3: The alternation learning problems

Similarly, the alternation learning problem can be formalized as follows: For input data S' (set of morphologically segmented words), and the hypothesis space $\mathcal{H}_{\text{Alternation}}$, the learning objective is to identify underlying representations $\mathcal{U}$ corresponding to S', then identify an alternation grammar $G_{\text{Alternation}}$ from the hypothesis space $\mathcal{H}_{\text{Alternation}}$ such that $G_{\text{Alternation}}$ approximates the target grammar $\mathcal{T}_{\text{Alternation}}$. $\mathcal{H}_{\text{Alternation}}$ represents all possible UR-SR mappings, and $\mathcal{T}_{\text{Alternation}} = \{f : \mathcal{U} \rightarrow \mathcal{L}\}$ is the set of functions that map the learned underlying representations to surface representations accepted by the target language $\mathcal{L}$.

To summarize, modeling phonological learning fundamentally involves understanding how humans derive a hypothesis grammar G for phonotactics and alternation from the input data S. These formal definitions illuminate the connection between phonotactics and alternation learning. In both phonotactic and alternation learning, the goal is to acquire a grammar that closely approximates the target language $\mathcal{L}$. As detailed below (§6.2), the grammar $G_{\text{Phonotactics}}$ learned from phonotactic learning assists in alternation learning by filtering out lexical exceptions that are not part of the target language $\mathcal{L}$ from the input data.

6.1.4 Phonotactics as an exception-filter in alternation learning

Building upon previous work, this proposal connects the link between phonotactics and alternation learning to the problem of EXCEPTIONALITY (Moore-Cantwell and Pater, 2016; Dai, 2023; Finley, 2010). The current study shows that, phonotactics can serve as a *filtering mechanism* of phonotactically illicit data that allows the learner to pay attention to phonotactically licit data, hence facilitating alternation learning.

This dissertation develops a two-stage phonotactic-alternation learning model (henceforth "Two-stage learner") that learns phonotactics and alternation from a shared input data. In the phonotactic learning stage, the learner is trained on morphologically unlabelled sound sequences of the input data and returns a set of categorical phonotactic constraints that penalize statistically underrepresented sound sequences. In the alternation learning stage, the learner obtains the access to morphological segmentations of the input data, and can utilize the learned phonotactic grammar to exclude phonotactically illicit sound sequences.

Crucially, when phonotactics and alternations are isomorphic, lexical exceptions that violate alternations are also phonotactically illicit. In cases of phonotactics-alternation mismatch, although lexical exceptions of phonotactics and observed alternations are unrelated, a successful phonotactic learning is still crucial for alternation learning, because an over-restricted phonotactic grammar will eliminate too many surface forms that are necessary for successful alternation learning. Consider the regular Turkish back harmony pattern mentioned above, after the back vowel [ɑ], the following vowel productively changes to [+back] across root-affix boundaries. However, lexical exceptions that violate this pattern, such as [sɑɑt-in] "hour-GEN" where [i] is [-back], can surface in child-directed speech (Slobin, 1982; MacWhinney, 2000; Belth, 2023a, Aksu and Altinkamis corpora within CHILDES database). The Two-stage learner will filter out phonotactically illicit sequences such as [ɑ…i] (on a vowel tier) during alternation learning, enabling featural generalizations such as /i/ → [+round]/[+round]_. Phonotactic restrictions that

are not associated with alternations will neither facilitate nor block alternation learning, such as the positional restriction of mid round vowels in Turkish.

When evaluated on unseen data, the proposed model excels in learning Turkish vowel harmony, surpassing previous models, including the Transformer-based seq-to-seq language model in Belth (2023b:Chapter 4). Moreover, the model successfully learned vowel harmony in real-world Finnish and Hungarian corpora, which consist of a large amount of lexical exceptions that don't undergo vowel harmony (Duncan, 2015; Szeredi, 2016).

The current study compared three different strategies in learning alternations, including minimal feature-based, maximal feature-based, and segment-based generalizations, as defined in §6.2. The results show that, in the Two-stage learner, minimal and maximal feature-based generalizations both outperform the segment-based generalizations, while maximal feature-based generalizations overall achieved the highest performance and is substantiated by recent experimental evidence (Pycha et al., 2003; Durvasula and Liter, 2020).

6.2 Proposal: Two-Stage Phonotactic-Alternation Learner

For the convenience of discussion, this dissertation uses Turkish vowel harmony as a working example to illustrate the proposed learning model. Turkish vowels are shown in Table 6.2.

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

Table 6.2: Turkish vowel system

The feature system designates the featural specifications of segments, exemplified in Table 6.3. Among all phonological features, those universally shared across segments on a given

tier are termed as *non-contrastive* features. These are excluded since they remain static during alternation learning. For example, the primary features under consideration for segments on the vowel tier are 'high', 'round', and 'back', and features such as 'syllabic' and 'labial'[4] are non-contrastive and consequently do not condition phonological alternations.The current study abstracts away from the problem of feature learning (Mayer, 2020; Dai et al., 2023). For the convenience of discussion, I employed conventional shorthand symbols for archiphonemes, also known as ARCHIPHONEME (Trubetzkoy, 1939, 1969). These symbols correspond to shared features of alternants in the current study. For instance, the symbol /H/ denotes the shared feature of all high vowels [ɯ, i, u, y], whereas /A/ represents all low unrounded vowels [ɑ, e]. Likewise, /I/ stands for all high front vowels [i, y].

Segments	**High**	**Back**	**Round**
i	+	-	-
u	+	+	+
y	+	-	+
ɯ	+	+	-
ɑ	-	+	-
e	-	-	-
o	-	+	+
ø	-	-	+
H	+	0	0
I	+	0	-
A	-	0	-

Table 6.3: Turkish feature system (vowels), omitting non-contrastive features

[4]This proposal operates on the assumption that vowels are assigned a 0 value for all PLACE features. To account for consonant-vowel interactions, such as labial attraction (Lees, 1966; Zimmer, 1969; Kabak, 2011), one would need to adjust the assumed tier and corresponding feature system.

Turkish vowel phonotactic patterns introduced in chapter 4 are summarized 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 non-initial syllable of a word, which means that they cannot follow other vowels.

Only the first two phonotactic patterns are associated with morphophonological alternations in stem + suffixes.

Generally, a substantial number of exceptions to vowel harmony patterns arise from compounds and loanwords (Lewis, 2001; Göksel and Kerslake, 2004; Kabak, 2011). For example, the compound word [bugyn] "today" ([bu] "this" + [gyn] "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. Moreover, segments certain morphemes neither undergo nor block vowel harmony, such as /i/ in /-Abil/ and /-ki/.

Despite many exceptions, these generalizations are not only well-documented in the literature, including Underhill (1976:25), Lewis (2001:16), Göksel and 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 in head-turning paradigm experiments. (Altan et al., 2016). As mentioned in chapter 4, Turkish vowel phonotactic constraints are applicable within roots and across morpheme boundaries (Zimmer, 1969; Arik, 2015). 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.

To summarize, Turkish vowel harmony presents a unique challenge for phonological learning: How does the learner acquire the alternations in vowel harmony patterns that predict phonotactically licit SRs, despite the high level of lexical exceptions within the input data? This challenge is tackled in the proposal below.

6.2.1 Overview:

The two-stage phonotactic-alternation learner (henceforth "Two-stage learner") models how children acquire phonotactic and alternation grammars from real-world data, as formalized above. The input data of the learner includes the training data S, tier, feature system F, and the threshold parameter for phonotactic learning $\theta_{\max}$. Table 6.4 shows an example of the training data, which includes surface representations (SR), the morphologically parse SRs and corresponding glosses.

SR	Parsed SR	Gloss
dynjɑɯn	dynjɑ-ɯn	world-GEN
kiʃiin	kiʃi-in	person-GEN
kɑnunun	kɑnun-un	law-GEN
gynyn	gyn-yn	day-GEN

Table 6.4: A sample of training data for alternation learning from Turkish

The learner undergoes two learning stages, as shown in Figure 6.4. In the phonotactic learning stage, the training data consists of the morphologically agnostic surface representations (SRs). Using a phonotactic learning algorithm as detailed in Dai (2023), the learner formulates a categorical phonotactic grammar. Crucially, this learning is not limited to roots or specific morphological contexts. The outcome of this stage is a set of categorical phonotactic constraints, such as {*ɑe, *ɑi, *øi, ...}.

In the alternation learning stage, the learner gains access to both SRs and morphological segmentations. This stage's training data encompasses morphologically parsed SR accompanied by their labels (or gloss). Table 6.4 showcases a sample of the Turkish training data. The underlying representations (URs) are initialized following a set of principles, which will be elaborated below. Given the morphological paradigm and initialized URs, the learner then transitions to the alternation learning module, and output a set of phonological alternations expressed as phonological rules of UR-SR mappings (Chomsky and Halle, 1968; Albright and Hayes, 2003b).

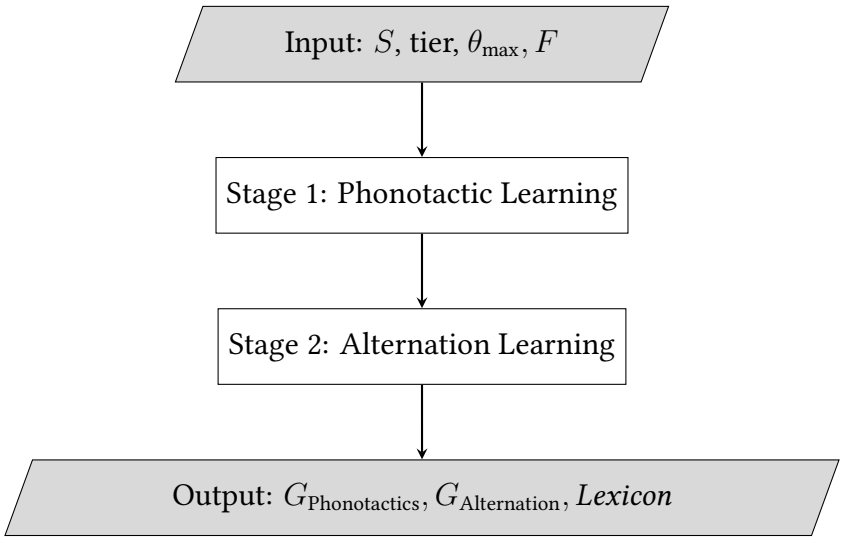


Figure 6.4: Two-Stage Phonotactic-Alternation Learner

The innovative proposal in the current study, at a COMPUTATIONAL LEVEL (Marr, 1982), is that phonotactics functions as a filtering mechanism of phonotactically illicit input data for alternation learning. This filtering mechanism is substantiated by experimental evidence, e.g., Lentz and Kager (2015) showed the role of L1 categorical phonotactic knowledge in filtering L2 input. In the ALGORITHMIC LEVEL, there are several working hypotheses employed to facilitate the evaluation against real-world corpora. In particular, the current study employs discrete, categorical grammars for both phonotactics and alternations, and abstracts away from the problem of gradient productivity in phonological processes (Albright and Hayes, 2003a; Hayes et al., 2009;

Moore-Cantwell and Pater, 2016).[5]

Moreover, learned alternations are described in a rule-based formalism as in (Albright and Hayes, 2003a). Extending the Two-stage learner to constraint-based grammars requires non-trivial modification to ensure a finite candidate set of possible SRs for a given UR, as discussed in chapter 8. Moreover, rule-based UR-SR mappings serve as a mathematical characterization that underlies any computational model of phonological acquisition, including constraint-based frameworks.

6.2.2 Stage 1: Phonotactic Learning

The phonotactic learning algorithm follows Dai (2023)'s exception-filtering phonotactic learner. Built upon Hayes and Wilson (2008)'s approach, Dai (2023)'s observed-to-expected ratio (O/E) estimation updates the expected frequency (E) while concurrently filtering out lexical exceptions in the input data, which in turn updates the observed frequency (O). 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 θ_{max}, also known as the *accuracy schedule* (Hayes and Wilson, 2008), where the interval after 0.1 is fixed to 0.1. For example, if $\theta_{\text{max}} = 1$, then the accuracy schedule $\Theta = [0.001, 0.01, 0.1, 0.2, 0.3, \ldots, 1]$. This structure prioritizes the integration of potential constraints with the lowest O/E values. θ_{max} can be interpreted as follows: the higher θ_{max} indicates the need for more statistical support, i.e. higher O/E, before considering a two-factor as grammatical. There is a tolerance for a certain degree of observed frequency being lower than expected in the majority of real-world datasets analyzed, and the appropriate θ_{max} value is usually below 0.7 (Dai, 2023).

As shown in Figure 6.5, given the input data S (morphologically unsegmented surface representations), tier, and the maximum O/E threshold θ_{max}, the learner first initializes an empty hypothesis grammar $G_{\text{Phonotactics}}$ and hypothesis space Con (Step 1). The learner then selects the

[5]See detailed discussion in Dai (2023:§2).

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 $G_{\text{Phonotactics}}$, 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.[6]The output of the phonotactic learning is set of TSL_2 constraints such as { *ui, *øi, *ɑi, ...} for the vowel tier in Turkish.

[6]Adriaans and Kager (2010) suggested a similar model that employs the classic O/E equation, originally introduced by Pierrehumbert (1993); Frisch et al. (2004), for selecting all 2-factors with an O/E ratio below a certain threshold simultaneously, without the iteration as in Dai (2023). However, Pierrehumbert (1993)'s equation presupposes an empty hypothesis grammar, a premise that loses accuracy with the addition of any constraints. This limitation has been discussed in Wilson and Obdeyn (2009) and Wilson (2022).

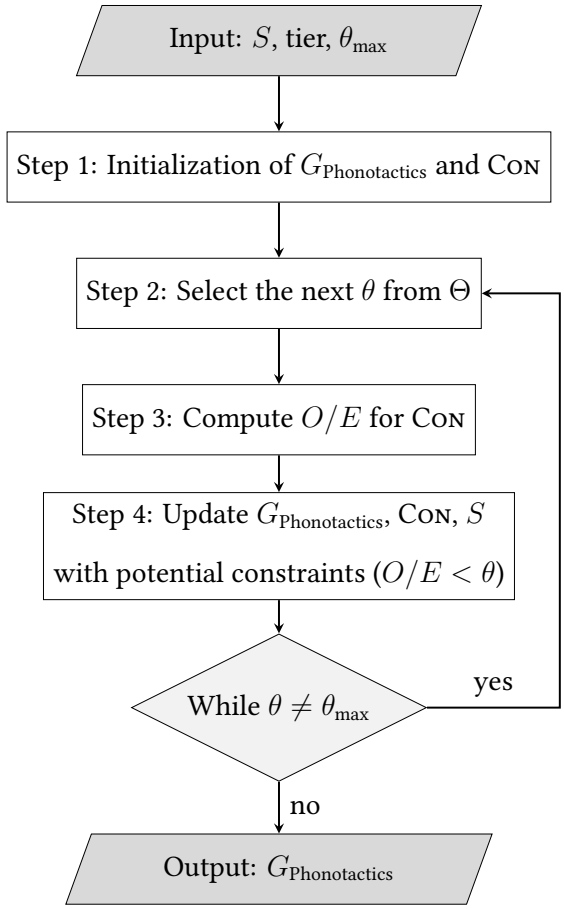


Figure 6.5: Phonotactic learning

6.2.3 Stage 2: Alternation learning

The input of alternation learning includes the input data (S), previously learned phonotactic grammar ($G_{\text{Phonotactics}}$), and the feature system (F). The input data consists of the morphologically parsed SRs. The role of morphological information is to provide the derived environment where alternations happen and provide the learner an access to URs stored in the lexicon. Given the input, the goal of alternation learning is to arrive at a phonological grammar which consists of rules that characterize the derivation from UR to SR in the observed data.

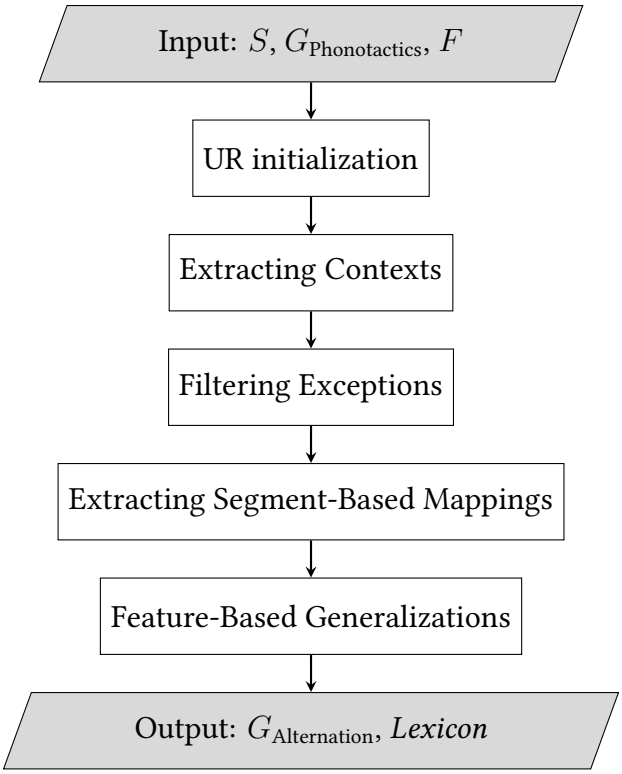


Figure 6.6: Alternation learning

UR Initialization. The current study assumes that UR learning forms an integral part of the construction of lexicon rather than the phonological grammar. Providing a comprehensive UR learning algorithm is beyond the scope of this dissertation on grammar learning. For recent reviews of the literature on this subject, one may refer to (Wang and Hayes, 2022; Belth, 2023a). Nevertheless, the current study utilized several guiding principles for UR initialization:

1. By default, the URs of stems and non-alternating forms are identical with their SRs.

2. For alternating segments on the prespecified tier in allomorphs, use their shared features to initialize abstract URs.[7] For example, the alternating vowels in allomorphs of Turkish GENITIVE [-in, -un, -ɯn, -yn] share [+high], which results in an abstract /H/ in the UR /-Hn/.

[7]The tier eliminates non-tier segments that can surface as a default SR in phonological processes, such as the neutral vowel [e] in Hungarian vowel harmony.

The output of the UR initialization is a dictionary of morpheme-UR pairs, e.g., {ACCUSATIVE: /H/, GENITIVE: /Hn/, ...} in Turkish.

Throughout the alternation learning component, allomorphs involving deletion and insertion are both omitted from the input data based on their distinct lengths and missing/additional segments. For example, in the allomorphs of Hungarian PRESENT.3SG: [-m, -em, -om, -am, -öm], -m lacks the shared vowel present in all other allomorphs and is shorter than all other forms in length. Therefore, it is omitted during UR initialization. For more thorough solutions on deletion and epenthesis, refer to Nyman and Tesar (2019) and Wang and Hayes (2022).

Again, this is a simplified mechanism for positing UR, which biases towards abstract URs of alternating segments and does not accommodate intricate cases such as segmental deletion or insertion. However, the posited URs serve as an initial hypothesis which can be improved by more sophisticated UR update algorithm in future work.

Extracting Contexts. The learner scans through all data, extracting alternating segments, corresponding URs, and preceding contexts in all allomorphs, as shown in Table 6.5.[8] Following the linguistic convention, the symbol _ represents the alternating segment. For example, e_ indicates the combination of an alternating segment and its preceding context /e/.

The context extraction process is restricted to tier-based derived environments. At the beginning of the procedure, the algorithm identifies the last tier segment in the stem, referring to it as the preceding environment. This extraction can be efficiently implemented by scanning the stem until reaching the first morpheme boundary. It is worth noting that, the extraction of stem environment is also used in rule application. The preceding environment is then continually updated as the learner progresses and examines subsequent SR segments. Extracting SR instead of UR contexts is determined by the structure of Output Tier-based Strictly Local functions. The complexity of this operation for each n-length string is $O(n)$.

[8]While it's possible to modify the learner to extract phonological contexts of individual morphemes for the purpose of learning *phonologically-conditioned morphology*, as per Nevins (2011), such an approach is not pursued in the current study focusing on phonological alternation.

Consider the UR-SR mapping /firmɑ$_1$-lA$_2$r-H$_3$n/→[firmɑ$_1$-lɑ$_2$r-ɯ$_3$n] as an example. The learner will initially set the stem-final [a$_1$] as the *preceding environment* for [a$_2$]. Subsequently, when the learner encounters ɯ$_3$, the *preceding environment* is updated to [a$_2$].

Filtering Exceptions. The learner then filters out phonotactically illicit sequences from context-segment pairs. For example, [øi] and [ɑi] in Table 6.5 can be eliminated by the learned phonotactics {*øi, *ɑi} from the filtered context. In other words, the learner uses phonotactics to filter out a large amount of *nondeterminism* during the alternation learning. In the absence of a successful phonotactic filter, it would be impossible to propose deterministic mapping to a unique SR for every distinct UR and context. For example, in the unfiltered contexts, /H/ becomes either [i] or [ɯ] following ɑ_, either [i] or [y] following ø_, and either [u] or [ɯ] following u_ and o_. [9] See more discussion on this approach in chapter 8.

Segments	Unfiltered	Filtered
i	[e_, i_, ø_, ɑ_]	[e_, i_]
u	[u_, o_]	[u_, o_]
y	[y_, ø_]	[y_, ø_]
ɯ	[ɑ_, ɯ_, u_, o_]	[ɑ_, ɯ_]

Table 6.5: Alternating segments for UR /H/ and preceding contexts, comparing unfiltered and exception-filtered data from the Aksu and Altinkamis corpora within the CHILDES database (Slobin, 1982; MacWhinney, 2000; Belth, 2023a)

In the meantime, the identified phonotactically illicit forms are stored in the *Lexicon*. The current study assumes the dual mechanism model (Pinker and Prince, 1988; Zuraw, 2000; Zuraw et al., 2021), in which lexicon and grammar are two distinct components of linguistic competence.

[9] An alternative approach for handling nondeterminism is by using *stochastic rules*, which assign different probabilities to co-existing phonological processes based on their statistical support (Albright and Hayes, 2003b; Gorman and Reiss, 2023). For example, hypothetically, following ɑ_, /H/ has 15% chance to become [i], and 85% chance to become [ɯ]. Moreover, this approach might be advantageous in explaining regularity in exceptional patterns, such as semi-harmonic context-segment pairs in Turkish that agree in [back] but differ in [round], e.g., øi, ɑi, and oɯ. This could be explained by low probabilities in the stochastic rule governing backness harmony.

During morpho-phonological processing, speakers initially consult observed forms in the lexicon. If encountering a nonce form, as in the wug test, then they resort to the acquired grammar.

Extracting Segment-Based Mappings. The learner extracts segment-based mappings by examining the UR to SR transitions of alternating segments across morphological paradigms in the filtered input data. Consider the filtered data in Table 6.5, the learner acquires mappings such as H → i / e_. The extraction of segment-based mappings lay the groundwork for the subsequent feature-based generalizations.

Feature-Based Generalizations. To facilitate the discussion of the featural generalizations, the contexts of segment-based rules are combined if they share a structural change. For example, H → i / e_ and H → i / i_ can be merged to H → i / [e, i]___. The learner infers feature-based generalizations from segment-based mappings by interpreting the structural change as a featural change from UR to SR. For example, for the UR /H/ ([+high]), [-back, -round] represents a change to [i] ([+high, -back, -round]). Moreover, the learner extracts feature-based representations of contexts, for example, [e, i]___ can be represented as [-back, -round]. [10] A future direction is to examine whether these unnatural classes observed in fieldwork data can be simplified by phonotactic filters during alternation learning.

At this step, there are two possible approaches for the learner to infer the specific feature-based mappings. The *Minimal Generalization* approach aims to induce the *most specific* featural descriptions, hence minimal generalization, based on the shared feature-based representation of structural changes and contexts (Albright and Hayes, 2003b). Take H → i/[e, i]___ as an example. The context [e, i] can be described as the exact set of shared feature-based representation [-back, -round], as illustrated in Table 6.6.

[10]Following Albright and Hayes (2003b), the current study assumes that it is possible to extract feature-based mappings and abstracts away from the problem of unnatural classes that might also condition phonological processes. For example, Kolami plural /-(u)l/ allomorphy (Emeneau, 1961; Mielke, 2008) is conditioned by the context of unnatural class [p, t̪, k, d̪, g, s, v, z, m, ŋ, j]___.

Segment-Based	Minimal Generalization
H → i / [e, i]___	H → [-back, -round] / [-back, -round]___
H → u / [u, o]___	H → [+back, +round] / [+back, +round]___
H → y / [y, ø]___	H → [-back, +round] / [-back, +round]___
H → ɯ / [ɑ, ɯ]___	H → [+back, -round] / [+back, -round]___

Table 6.6: Feature-based rules based on the minimal generalization approach

The current study also proposes an alternative *Maximal Generalization* approach which first deduces the most general one-feature rules, hence maximal generalizations, that does not contradict any segment-based mappings, as illustrated in Table 6.7. This approach begins with a hypothesis on feature-based phonological mappings, e.g., H → [-back]/ [-back] ___, which are then tested against observed segment-based mappings. Any contradiction between the segment-based mappings and a hypothesized feature-based mapping results in the latter's removal. For instance, encountering H → i / ɑ___ in segment-based mappings will lead to the exclusion of H → [-back] / [-back] ___ from the hypothesis grammar due to the contradictory [+back] feature of ɑ. At the end of the alternation learning, the learner adds segment-based rules that cannot be described by any valid hypothesized feature-based generalizations to the hypothesis grammar. This method finds its roots in classic linguistic analysis (Halle, 1961; Chomsky and Halle, 1968; Pinker and Prince, 1988), and is supported by recent experimental study on *Simplicity Bias* (Pycha et al., 2003; Durvasula and Liter, 2020, Multiple Simpliest Generalizations).

Segment-Based	Maximal Generalization
H → i / [e, i]___	H → [-back] / [-back]___, H → [-round] / [-round]___
H → u / [u, o]___	H → [+back] / [+back]___, H → [+round] / [+round]___
H → y / [y, ø]___	H → [-back] / [-back]___, H → [+round] / [+round]___
H → ɯ / [ɑ, ɯ]___	H → [+back] / [+back]___, H → [-round] / [-round]___

Table 6.7: Feature-based rules based on the maximal generalization bias

Although they often result in extensionally equivalent mappings, the maximal generalization approach is more advantageous in dealing with ACCIDENTAL GAPS in the input data. When a crucial context for the best-performing minimal generalization is missing, maximal approach can still hypothesize robust, general, featural generalizations. Consider the minimal generalization H → [+back, -round] / [+back, -round]___, which captures the segment-based mapping H → ɯ / [ɑ, ɯ]___. If all SRs with [ɑɯ] is accidentally missing from the input data, resulting in H → ɯ / [ɯ]___, then the minimal generalization will become very specific H → [+high, +back, -round] / [+high, +back, -round]___ and fail to predict H → ɯ / ɑ___ in the unseen data. Moreover, when more complex featural representations are involved in phonological processes, maximal generalization approach can drastically reduce the amount of rules.

The current study assumes NATURALNESS BIAS within the maximal generalization approach to maintain consistency in the feature domain across both structural change and context. Consequently, this approach avoids the incongruous generalizations which fail to capture natural assimilatory/dissimilatory generalizations internalized by native speakers such as H → [-round]/[-back]___. In Optimality Theory, this naturalness bias is often implicitly captured by feature-based constraints such as AGREE[F], OCP[F], and IDENT[F], which evaluate candidates within the same domain F.[11] There has been evidence for such bias from the angle of typological force that

[11]Although certain features recurrently pattern together, it is possible to group these features in a language-specific feature class (or *Color*). For example, Padgett (2002) grouped [back] and [round] in the same feature class

favors assimilation or dissimilation (Blevins, 2004).

as the domain of Turkish vowel harmony. Exploring potential representational system is beyond the scope of the current study.

CHAPTER 7

EVALUATION OF THE TWO-STAGE PHONOTACTIC-ALTERNATION LEARNING MODEL

In the following sections, the proposed Two-stage learner is evaluated on real-word corpora in Turkish and Finnish.

7.1 Datasets

Table 7.1 provides a sample of these datasets, including token frequencies and morphological segmentations.

SR	Parsed SR	Gloss	Token Frequency
dynjɑɯn	dynjɑ-ɯn	world-GEN	2,712
kiʃiin	kiʃi-in	person-GEN	2,572
kɑnunun	kɑnun-un	law-GEN	1,579
gynyn	gyn-yn	day-GEN	941

Table 7.1: A sample of training data for alternation learning from Turkish

In each case study, datasets are partitioned into training and testing sets based on the ten-fold split, as shown in Figure 7.1.

During the training phase, the learning model is exposed to both stems and derived forms (stem + affixes). However, in the testing phase, evaluation focuses exclusively on predicting alternations in tier segments and stem-affixes combinations when calculating the predicative accuracy. Notably, the current study excludes consideration of alternations involving nontier segments, such as t/d voicing (Belth, 2023a), and insertion/deletion processes, as these fall outside the scope of this dissertation.

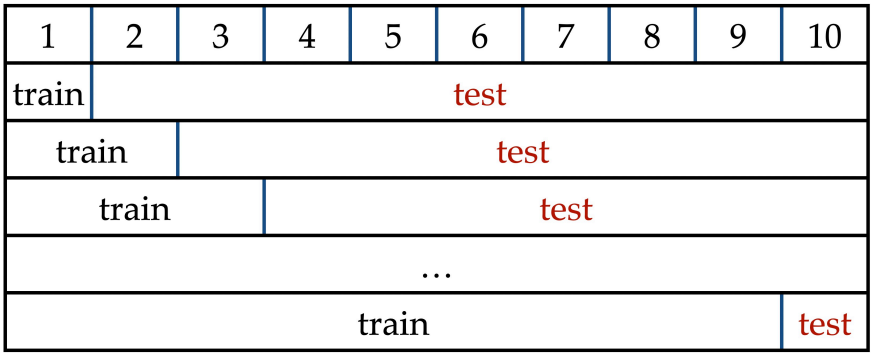


Figure 7.1: Ten-fold split

Following Belth (2023b), the training employs a batch learning method, beginning with an initial subset (10% of the data) and incrementally incorporating the entire dataset, with batches sampled based on token frequency with replacement. In other words, the learning model is first exposed to the first 10% of the dataset and tested on the rest 90%, then trained on the first 20% and tested on the rest 90%, and so on. This approach simulates natural language acquisition, prioritizing exposure to high-frequency words before introducing less common terms. Moreover, this method does not process learning data one by one as in online learning (Tesar and Smolensky, 2000).

As detailed in §6.2, the learning model first acquires a segment-based categorical phonotactic grammar. Once morphological segmentation is introduced, the learner begins to infer feature-based morphophonological alternations.

7.1.1 Evaluating Phonotactic Learning

To interpret the result of phonotactic learning, I created a test dataset of 64 nonce words following the $[CT_1CT_2C]$ for every case studies, where C represents non-tier consonants and T represents tier segments, e.g., [tokuz]. Every nonce word corresponds to one possible TSL_2 factor ($T_1...T_2$) to avoid any sampling bias that might arise from manually reducing or increasing the amount of certain combinations. Only roots are included, as the phonotactic learning in the Two-stage

learner disregards morpheme boundaries.

Each word is categorically labelled 1 ("grammatical") or 0 ("ungrammatical") based on converging evidence from the well-documented phonotactic generalizations and behavioral experiments. In particular, the grammaticality labels for the Turkish are supported by behavioral experiment (Zimmer, 1969), as discussed in Dai (2023).

This follows the standard practice in previous computational studies when acceptability judgments of nonce words in the test dataset are not accessible. For example, Gouskova and Gallagher (2020) manually labelled the categorical grammaticality of nonce words in the test dataset based on documented phonotactic generalizations supported by behavioral experiments (Gallagher, 2014, 2015, 2016).

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 any types of errors yields an F-score of 1.

7.1.2 Evaluating Alternation Learning

After each training session, the model's performance is tested on the remaining, unseen dataset (the test dataset). Given the gloss and induced URs, the learned alternation grammar predicts corresponding SRs. The predicted SRs are compared with observed SRs, and the accuracy is the number of correct predictions divided by the size of test dataset.[1] A successful learning model should achieve a high accuracy on these test datasets, effectively predicting the surface forms of derived stem + affix combinations.

[1]It is worth noting that smaller training batches result in larger test datasets, increasing the statistical power of the model. This means a mistake made in a small batch can cause a larger penalty in performance than a large batch, as shown in the dip at 5k words of Turkish test result ($\theta_{\max} = 0.4$; MorphoChallenge) in Figure fig. 7.4 below.

When evaluating the alternation learning, three approaches are compared: (1) a model with segment-based mappings (henceforth Segment-based); (2) tightest fit minimal feature-based generalization (henceforth Minimal) (3) maximal generalization with naturalness bias (henceforth Maximal). In the learned segment-based mappings, conflicting mappings like H→i/ɑ_ and H→ɯ/ɑ_ are excluded during evaluation to prevent ambiguity and ensure accuracy.

The current proposal predicts that, when the phonotactics filter fails to eliminate the exceptions from the contexts of segment-based mappings, the learner has difficulty learning the feature-based rules from the dataset. When the phonotactic grammar is too restrictive, realized as a high $\theta_{\max}$, treating too many observed sequences ungrammatical, the learner also fails in generalizing morphophonological alternations or predicting the unseen data. This is consistent with the acquisition evidence for Tolerance Principle (Yang, 2016), which argues that there is a bound in the memorization of lexical exceptions in the lexicon. When $\theta_{\max}$ is set too high, the learner essentially stores too many input data as lexical exceptions to the lexicon.

7.2 Case study: Turkish vowel harmony

This section applies the learning model to vowel harmony in real-world Turkish corpora.

7.2.1 Turkish Data

The representational system (features and archiphonemes) utilized for the Turkish data is detailed in chapter 6. The evaluation employed datasets from Turkish MorphoChallenge (Kurimo et al., 2010) and Aksu (2;0-4;8) and Altinkamis (1;4-2;4) corpora within the Turkish CHILDES database (Slobin, 1982; MacWhinney, 2000; Belth, 2023a). I used Belth (2023a) data, where low token frequency occurrences and those with ambiguous morphology ($\approx$ 1000) are removed from the corpora. The MorphoChallenge dataset contains 22,315 morphologically-segmented words. The CHILDES dataset, representing child-directed speech (parentese), contributes a 1,727 morphologically-segmented words. Table 7.2 shows the type frequency of two-factors in both

datasets, and two-factors that follow the aforementioned nonlocal vowel phonotactics are highlighted. CHILDES data includes fewer attested disharmonic two-factors than in MorphoChallenge.

	i	e	y	ø	ɯ	ɑ	u	o
i	19	65	5	0	0	51	0	16
e	276	264	2	0	0	49	4	4
y	6	48	44	0	0	6	2	0
ø	0	46	29	3	0	0	0	0
ɯ	1	2	0	0	66	91	0	0
ɑ	76	95	0	4	378	502	39	28
u	1	6	4	0	0	27	54	1
o	8	4	5	0	0	13	93	29

(a) CHILDES

	i	e	y	ø	ɯ	ɑ	u	o
i	2,344	3,363	53	17	1	1,289	19	307
e	5,183	4,152	100	61	1	873	89	326
y	108	1,232	645	11	13	194	32	22
ø	14	817	444	4	0	7	0	12
ɯ	6	23	1	3	1,360	1,619	0	2
ɑ	1,537	1,284	80	61	4730	5116	315	473
u	67	106	22	1	3	1,600	966	23
o	440	362	69	12	3	1,274	1,031	347

(b) MorphoChallenge

Table 7.2: The type frequency of two-factors in the training datasets; highlighted cells correspond to two-factors allowed by the ideal phonotactic grammar.

7.2.2 Evaluating Learned Phonotactic Grammars

Figure 7.2 shows the blick test results on the on the full datasets of CHILDES and MorphoChallenge. The peak performance of the entire both datasets are at $\theta_{\max} = 0.4$ where the F-score is 0.9375 for CHILDES, and 1 for MorphoChallenge. This difference is due to different distributions of type frequency of vowel two-factors. Sampling is not employed to ensure the reproducibility of the plots, although similar outcomes are observed when incrementally sampling from the dataset with respect to the token frequency.

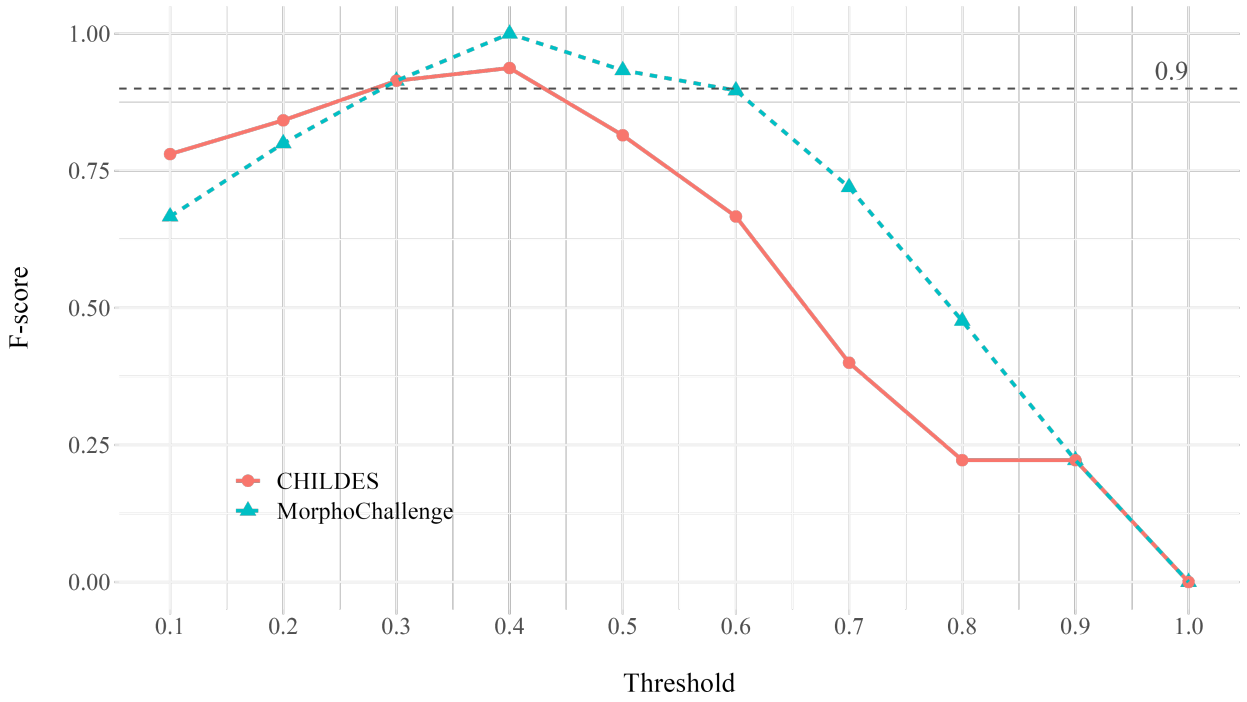


Figure 7.2: Turkish blick test result on the entire CHILDES and MorphoChallenge dataset; x-axis corresponds to θ_{max} values from 0.1 to 1; y-axis corresponds to the F-score in the blick test. The dotted line at y = 0.9 provides a visual guide of where the model performance peaks.

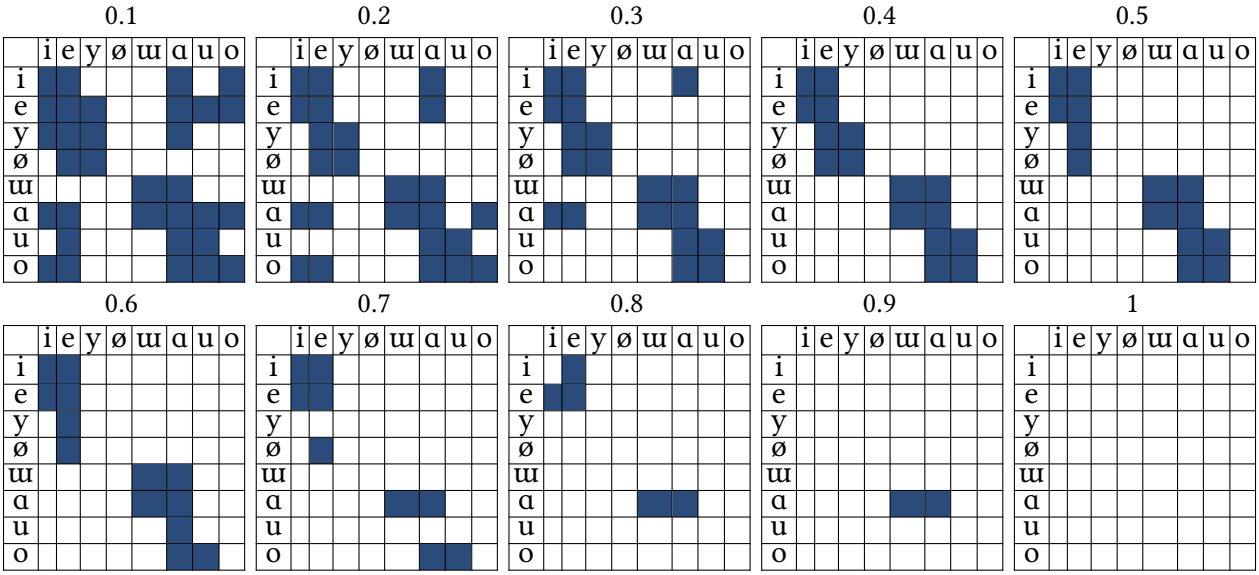


Figure 7.3: Learned phonotactics with the MorphoChallenge. Each individual subplot corresponds to a distinct $\theta_{\max}$ value, from 0.1 to 1, in phonotactic learning. Shaded cells indicates vowel two-factors accepted by the learned phonotactic grammar.

7.2.3 Evaluating Learned Alternation Grammars

After phonotactic learning, the Two-Stage learner filters out any morphologically parsed SRs that violate the learned phonotactic constraints from the input data. URs are initialized based on alternating segments in the entire dataset (as per UR Initialization in §6.2).[2]

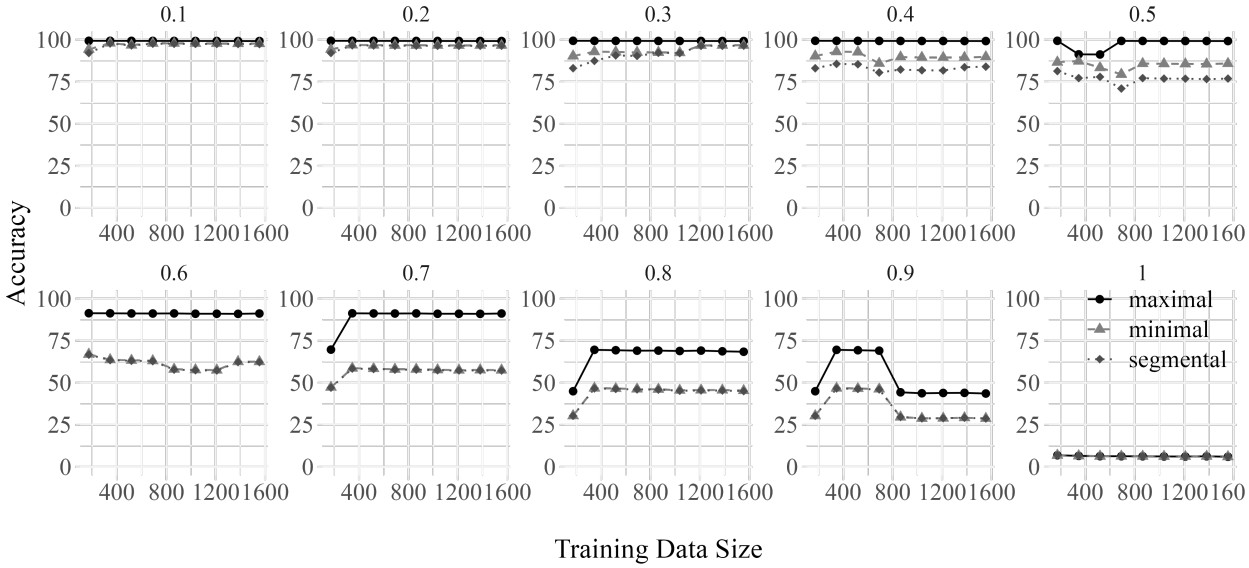
The learned alternation grammar is applied to the morphological gloss in the testing data, such as "world-GEN", and predict SRs. The predicted SRs are then compared to the real SRs in the test data. The accuracy is measured by (correct prediction)/(test dataset size). Figure 7.4 shows the learning trajectories of all models within the framework of Two-Stage learner. The learning result shows a comparable learning results in both minimal and maximal generalization, both outperforms segment-based generalizations in predicting unseen data.

During various trials, the Two-Stage (maximal) learner with $\theta_{\max} = 0.1$ can achieve highest accuracy scores when training in <3k-words dataset. When evaluating against the CHILDES data,

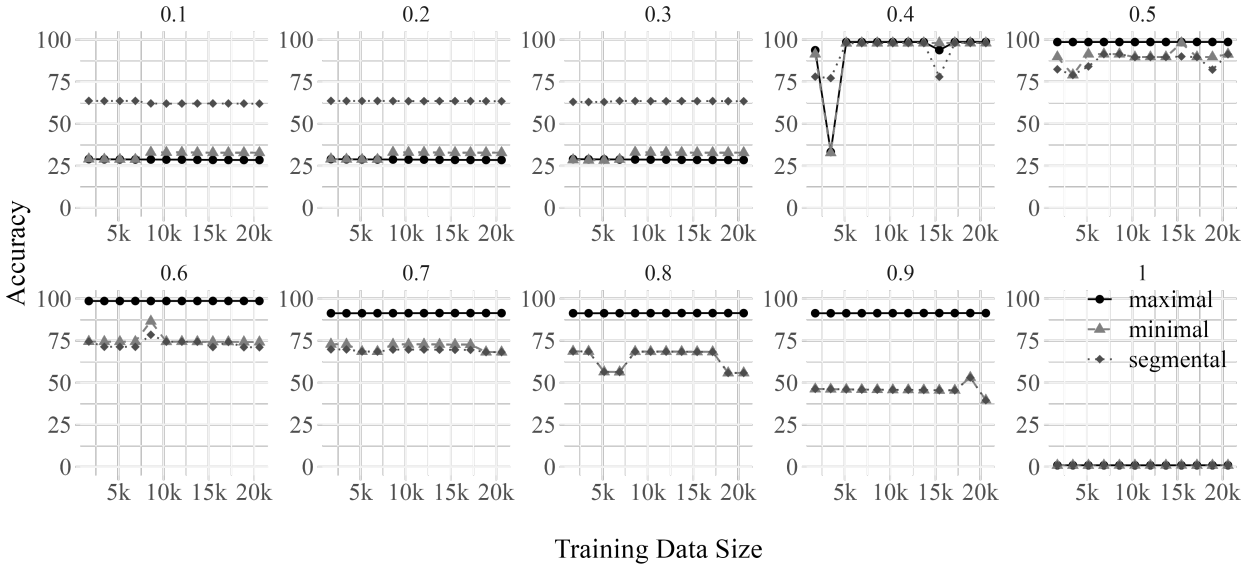
[2]The simplified UR learning facilitates the error analysis, otherwise errors in the evaluation might come from either URs or alternations. A comprehensive UR learning algorithm is beyond the scope of the current dissertation.

phonotactic learning seems redundant at the beginning—a phonotactic grammars at $\theta_{\text{max}} = 0.1$ that under-penalize several disharmonic two-factors is sufficient for learning alternations. This is because most disharmonic patterns in the derived environment allowed by such phonotactic grammars do not exist in this small dataset. In turn, the learner can still achieve robust feature-based generalizations despite not acquiring phonological knowledge regarding disharmonic patterns.

When learning from the larger MorphoChallenge dataset, however, a successful phonotactic learning becomes crucial. The phonotactic grammar that achieves a higher score in the blick test, especially when $\theta_{\text{max}} = 0.4$, results in overall better performance in alternation learning. If the threshold is set too low, the phonotactic grammar will allow too many exceptional patterns, while a too high threshold penalize too many sequences—both hinders the alternation learning. This can be observed by comparing the learned phonotactic grammars in Figure 7.3 and alternation grammars in Figure 7.4 at each threshold.



(a) Turkish test result in CHILDES



(b) Turkish test result in MorphoChallenge

Figure 7.4: Turkish test results in different test dataset; each individual subplot in (a) and (b) corresponds to a distinct $\theta_{\max}$ value, from 0.1 to 1, in phonotactic learning. The x-axis corresponds to the size of the training data, and the y-axis corresponds to the accuracy in predicting test data. Maximal and minimal generalizations have many overlaps.

Figure 7.5 compares the predicted accuracy of phonotactic and alternation learning. The

two-stage learner predicts that if the necessary phonotactic constraint is unlearned, the learner will fail to learn vowel harmony patterns. This can be monitored by adjusting the threshold level in the proposed phonotactic learner. For example, failing to acquire *a...e means the learner won't be able to learn the full backness harmony pattern. However, once this constraint is learned, the accuracy in predicting the novel forms spikes to almost 100%.

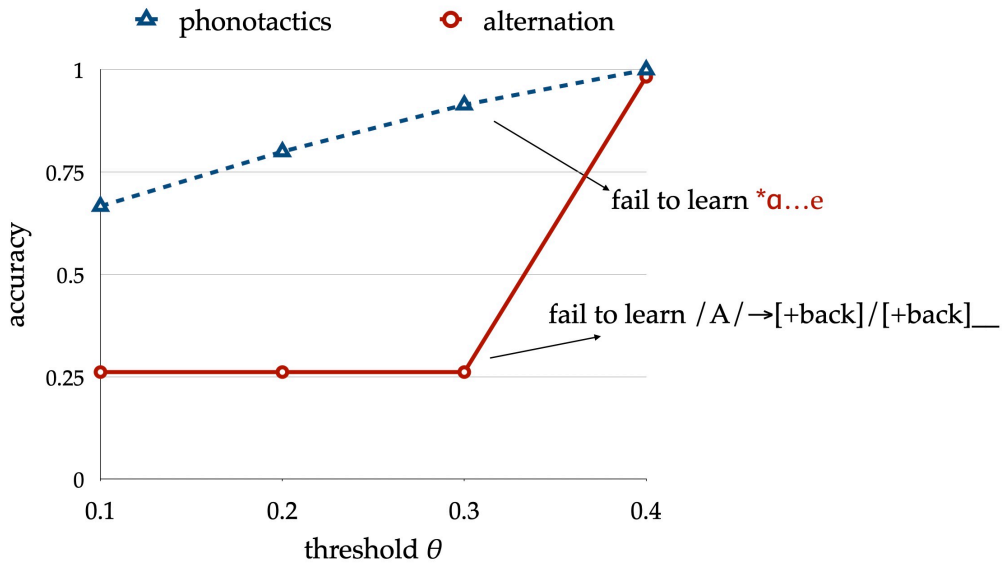


Figure 7.5: Comparing the accuracy of phonotactic and alternation learning

7.2.4 Error Analysis

After the training phase, the learner acquired similar learned alternation grammar $G_{\text{Alternation}}$ from full MorphoChallenge and CHILDES training data at $\theta_{\text{max}} = 0.4$, which is shown in Tables Table 7.3.

Segment-Based	Minimal Generalization	Maximal Generalization
H → i / [e, i]___	H → [-back, -round]/[-back, -round]___	H → [+back] / [+back] ___
H → y / [y, ø]___	H → [-back, +round]/[-back, +round]___	H → [-back] / [-back] ___
H → ɯ / [ɑ, ɯ]___	H → [+back, -round]/[+back, -round]___	H → [+round] / [+round] ___
H → u / [o, u]___	H → [+back, +round]/[+back, +round]___	H → [-round] / [-round] ___
I → ɯ / [ɯ, ɑ]___	I → [+back]/[+back,-round]___	I → [+back] / [+back] ___
I → i / [i, e]___	I → [-back]/[-back,-round]___	I → [-back] / [-back] ___
A → ɑ / [ɑ, ɯ, o, u]___	A → [+back]/[+back]___	A → [+back] / [+back] ___
A → e / [y, e, i, ø]___	A → [-back]/[-back]___	A → [-back] / [-back] ___

Table 7.3: Learned Turkish alternations by the Two-Stage Phonotactic-Alternation Learner ($\theta_{\max} = 0.4$)

This section analyzes the errors of maximal generalization in MorphoChallenge when $\theta_{\max} = 0.4$, which yields the highest accuracy in blick test of the phonotactic grammar. The maximal generalization exhibits a U-shaped learning trajectory from 93.74% ($\approx$ 1.7k words) to 33.28% ($\approx$ 3.3k words) accuracy, then rise back to 98.79% ($\approx$ 5k words) accuracy. This is because the phonotactic grammar failed to include *ɑe, which has a relatively high token frequency in the sampled 3.3k words. This constraint is crucial for eliminating the counter-evidence A → e / ɑ _. A → [+back] / [+back] _. Moreover, this failure results in the presence of e in the context of A → ɑ/ [y, e, i, ∅, ɑ], which impedes the learner's ability to generalize A → [-back] / [-back]_. The learner is able to learn correct alternations once the phonotactic grammar includes *ɑe at $\approx$ 5k words. After the first dip, the Two-Stage learner stay at 98.79% accuracy until a slight dip to 94.01% at 15k words due to its inability to include the *i in the phonotactic grammar, which contaminates the context of H→ i / [e, i, ɑ] _, resulting in the rejection of A → [-back] / [-back]_. These U-shaped learning trajectories shows the critical role of phonotactic learning in the Two-stage learner.

After training on 20k words, the learner reaches 98.82% (11759/11900) accuracy with 141

errors, all of which come from loanword stems. The combined size of training and test datasets is larger than the original size of MorphoChallenge, because the training dataset is sampled with replacement, thereby allowing repetition, whereas the test dataset comprises the remaining data. Table 7.4 lists these disharmonic stems and suffixes that caused errors in the evaluation. The gloss and origin is annotated based on an etymological dictionary *Nişanyan Sözlük* (Nişanyan, 2018).[3] Before the learner encounters these loanwords, they are treated as nonce words, resulting in the overregularization of vowel harmony that has been observed in Turkish-speaking children as young as two years old (Altan, 2009; Belth, 2023a). The special status of loanwords in phonological grammar has also been shown in Turkish stress assignment (Kabak and Vogel, 2001). One conjecture is that these exceptional derived forms from loanword are lexical exceptions memorized by native speakers in the lexicon, deviating from the productive phonological grammar.

[3]This dictionary is also available at https://www.nisanyansozluk.com/.

Stem	Disharmonic Suffixes	Gloss	Origin
gol-	-den	goal (in sports)	English
jar-	-i, -in	helper	Persian
sembol-	-ler, —y	symbol	French
sinjal-	-ler	signal	French
protokol	-ler, -yn, -y, -e, -de	protocol	French
petrol-	-y	petrol	French
materjal-	-ler, -i, -in	material	French
kristal-	-ler, -in	crystal	French
kontrol-	-yn, -den	control	French
santral-	-ler	central	French
anormal-	-lik	abnormal	French
festival-	-e	festival	French
final-	-e, -de	final	French
global-	-leʃ	global	French
alkol-	-yn, -e, -syz, -ly	alcohol	French
dikkat-	-miz, -niz, -ler, -li, -i	attention, care	Arabic
ʤemaat-	-in, -i	community	Arabic
gajrimenkul-	-ler, -yn	real estate	Arabic
gazal-	-i	gazelle	Arabic
hal-	-ler, -siz, -i, -in, -miz, -m-e	phase	Arabic
hajal-	-i	dream	Arabic
usul-	-de	method	Arabic

Table 7.4: Selected loanword stems and disharmonic suffixes that caused errors in the Two-Stage learner (maximal; $\theta_{\max} = 0.4$; 90/10 train-test split of MorphoChallenge data); all surface forms are transcribed in IPA; The UR of the vowel [e] in suffixes is /A/, while the UR for all other vowels in suffixes is /H/.

7.2.5 Cross-Framework Model Comparison

Figure 7.6 compares the Two-Stage learner with Belth (2023a)'s learner, which learns both URs and feature-based morphological alternations. Moreover, Belth (2023a) provided the learning result of a Transformer-based (Vaswani et al., 2017) seq2seq model, which is also included in the comparison. Belth (2023a) only reported the learning result on ≈3k words from MorphoChallenge dataset (weighted by token frequency), following the sampling with replacement method. The Two-Stage (maximal) learner with $\theta_{\text{max}} = 0.1$ achieves the highest accuracy score when training in <3k-words dataset, and is applied to both datasets here to facilitate the comparison with Belth (2023a).

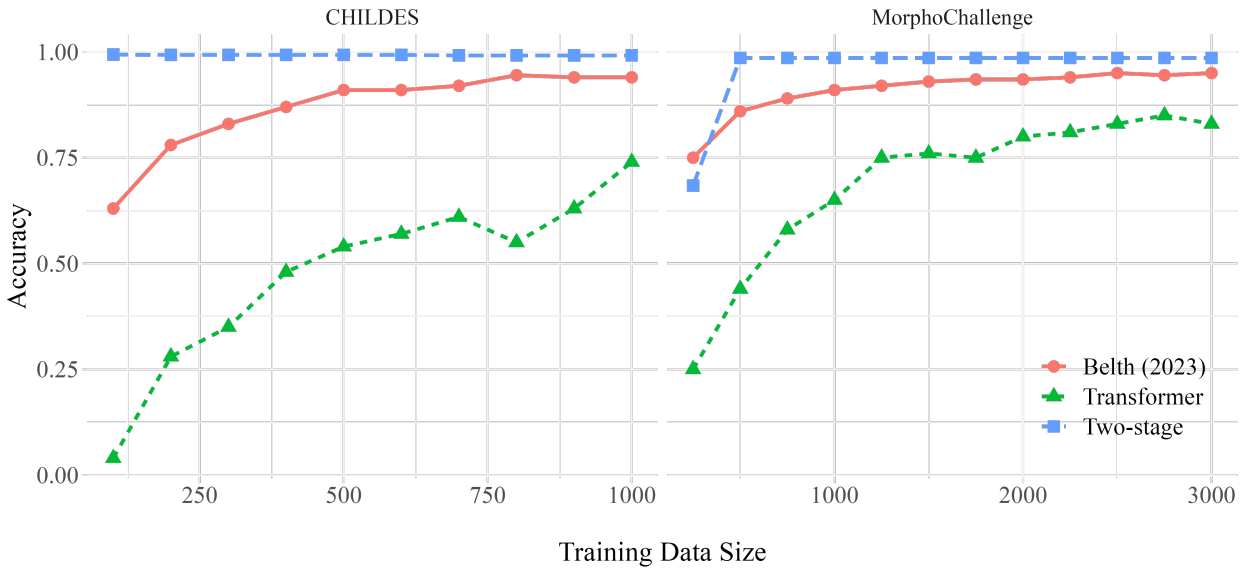


Figure 7.6: Model comparison among Belth (2023b), Transformer, and the Two-Stage model ($\theta_{\text{max}} = 0.1$; Maximal) in the current study; subplots correspond to distinct MorphoChallenge and CHILDES corpora; x-axis corresponds to the size of the training data; y-axis corresponds to the accuracy in predicting test data.

As shown in Figure 7.6, the current proposal can outperform the Transformer-based seq2seq model. However, its worth noting that Belth (2023a) also learns URs during the training, while the current study keeps constant URs initialized before alternation learning. Therefore, it is best

to conclude that Belth (2023a) and the current proposal provide comparable accuracy in both datasets.

7.2.6 Summary and Theoretical Implications

This case study on Turkish highlights two key findings: Firstly, the Two-Stage learner emerges as a highly effective model for mastering both phonotactics and alternations. Secondly, it underscores the pivotal role of phonotactic learning in conditioning alternation learning, especially vital in managing extensive datasets with numerous exceptional patterns like those in the MorphoChallenge data.

Despite varying type frequency distributions in the sampled training data, which could lead to various phonotactic generalizations, the proposed learner consistently and robustly acquires alternations, provided that key phonotactic constraints are initially learned.

7.3 Case study: Finnish vowel harmony

This section applies the Two-stage learner to Finnish vowel harmony.

7.3.1 Finnish Data

Finnish vowels are shown in Table 7.5. Long vowels are omitted, assuming their relevant feature representations follow their short vowel counterparts. Finnish orthography is converted to IPA, including *ö* [ø] and *æ* [æ].

	[-back]		[+back]	
	[-round]	[+round]	[-round]	[+round]
[+high, -low]	i *i*	y *y*	-	u *u*
[-high, +low]	e *e*	ø *ö*	-	o *o*
[-high, +low]	æ *ä*	-	ɑ *a*	-

Table 7.5: Finnish vowel chart; neutral vowels in vowel harmony are in shaded cells.

In Finnish, {i, e}, aka. NEUTRAL VOWELS (N), can co-occur with any vowels within a morpheme or across morpheme boundaries. However, non-neutral vowels are subject to a harmony pattern where back vowels (B) and front vowels (F) do not co-occur ($\neg$FB $\wedge$ $\neg$BF) either within a single morpheme or across morpheme boundaries. This can be described by a Tier-based Strictly 2-Local phonotactic grammar {*ɑæ, *aø, ..., *uy}, given the tier of non-neutral vowels.

This harmony pattern affects not only phonotactics but also conditions morphophonological alternations, particularly in the vowels of suffixes which alternate in accordance with the preceding vowels. For example, the vowel in plural partitive (PART.SG.) can surface as either back or front vowels:

Pattern	NOM.SG.	PART.SG.	gloss	Pattern	NOM.SG.	PART.SG.	gloss
a. F-F	pyy	pyy-tæ	“hazel grouse”	b. B-B	puu	puu-tɑ	“tree”
c. FN-F	tælli	tælli-æ	“wallop”	d. BN-B	tɑlli	tɑlli-ɑ	“stable”
e. N-F	pii	pii-tæ	“silicon”				

Table 7.6: Morphophonological alternations conditioned by Finnish vowel harmony pattern (Duncan, 2015).

This can be further formalized in following tier-based mappings across morpheme boundaries: let /A/ represents the UR of ɑ and æ [0back, -round, -high, +low], and /O/ switches between

o and ø [0back, +round, -high, -low], /U/ alternates between u and y [0back, +round, +high, -low].

1. A → [+back]/[+back]__, A → [−back]/[−back]__

2. O → [+back]/[+back]__, O → [−back]/[−back]__

3. U → [+back]/[+back]__, U → [−back]/[−back]__

When all preceding vowels are neutral, front vowels often surface as the default vowels, as shown in the N-F example (e) of Table 7.6.

Finnish vowel harmony often interacts with other local phonological processes. For instance, in (c) and (d) of Table 7.6, /t/ is deleted in the context of i_æ. Additionally, the alternating vowels in certain morphemes matches the immediately preceding vowel. This is observed in the third person singular morpheme /-V/, as seen in *netota-a* "to clear" and *kirjaile-e* "to embroider." Additionally, the vowel ɑ changes to o when it precedes a high front vowel or glide, such as [i] or [j]. For example, compare the singular *vamma-n* vs. plural *vammo-j-n* forms of [vɑmmɑ] "injury".

Numerous studies have demonstrated that Finnish vowel harmony is largely productive in native words, loanwords, and nonce words, including evidence from language games, production experiments, and acceptability ratings (Campbell, 1980; Vago, 1988; Mahonen, 2011; Duncan, 2015). Exceptions to Finnish vowel harmony are often observed in recent disharmonic loans and among younger speakers. For instance, the usage of PART.SG -tA in *asfaltti-tei-tä* ("some asphalt roads") contravenes the regular vowel harmony pattern. Additionally, variations in harmony patterns frequently occur when front rounded vowels y, ø [-back] serve as the triggering context, particularly with respect to the disharmonic loans associated with front rounded vowels (Ringen and Heinämäki, 1999; Mahonen, 2011, Chapter 1). Moreover, in a production experiment, Leiwo et al. (2006) shows that the speech of 196 Finnish children aged 2 years and 6 months seldom violates vowel harmony, and most errors can be attributed to production difficulties with front

rounded vowels. They indicate the correlation between the learning trajectories of the vowel categories and related harmony patterns: {y, ø} are the last vowels acquired by children (2;4) (Itkonen, 1977; Iivonen, 1993) with more production errors (Leiwo, 1977), and the harmony patterns associated with these front rounded vowels also show delayed acquisition and greater variation.

Taken together, Finnish children are exposed to a linguistic environment where vowel harmony is prominently represented, yet the harmony patterns related to less frequent front rounded vowels are often acquired later or may be completely lost as observed among some younger speakers (Mahonen 2011:87). This observation aligns with the simulation result of the proposed two-stage learner, as detailed in the following sections.

Similar to the Turkish case study, the evaluation employed the Finnish dataset from MorphoChallenge (Kurimo et al., 2010). The MorphoChallenge dataset contains 1,835 morphologically-segmented words. Table 7.7 shows the two-factors that follow the aforementioned nonlocal vowel phonotactics are highlighted. While attested, the vowel two-factors that characterize the front harmony patterns, such as ø…ø (4), are less frequent than those of back harmony patterns such as o…o (128). The low frequency of front rounded vowels and related patterns has also been observed in previous corpus studies (Pääkkönen, 1990:9; Szeredi, 2016:199).

	y	ø	æ	u	o	ɑ
y	39	49	49	10	13	16
ø	14	4	21	6	4	18
æ	40	27	108	13	4	17
u	17	0	4	169	227	300
o	11	1	5	182	128	281
ɑ	9	1	28	273	187	499

Table 7.7: The type frequency of two-factors in the training data; highlighted cells correspond to two-factors allowed by the ideal phonotactic grammar. Neutral vowels are omitted.

7.3.2 Evaluation Results

Figure 7.7 shows the blick test results on the on the full dataset of MorphoChallenge. The model achieves >0.9 performance when θ_{max} is between 0.2 and 0.5. Figure 7.8 shows the learned Finnish phonotactic grammar. Crucially, the phonotactic grammar eliminates all disharmonic two-factors at $\theta_{\text{max}} = 0.3$, while penalizing several underrepresented harmonic two-factors related to front vowels.

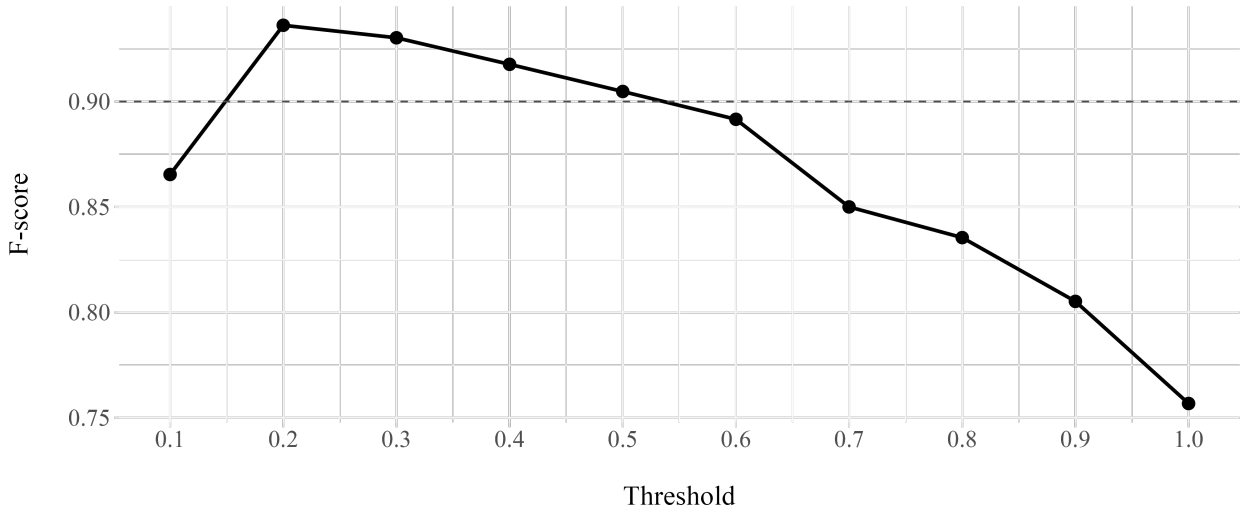


Figure 7.7: Finnish blick test result on the entire MorphoChallenge dataset; x-axis corresponds to θ_{max} values from 0.1 to 1; y-axis corresponds to the F-score in the blick test.

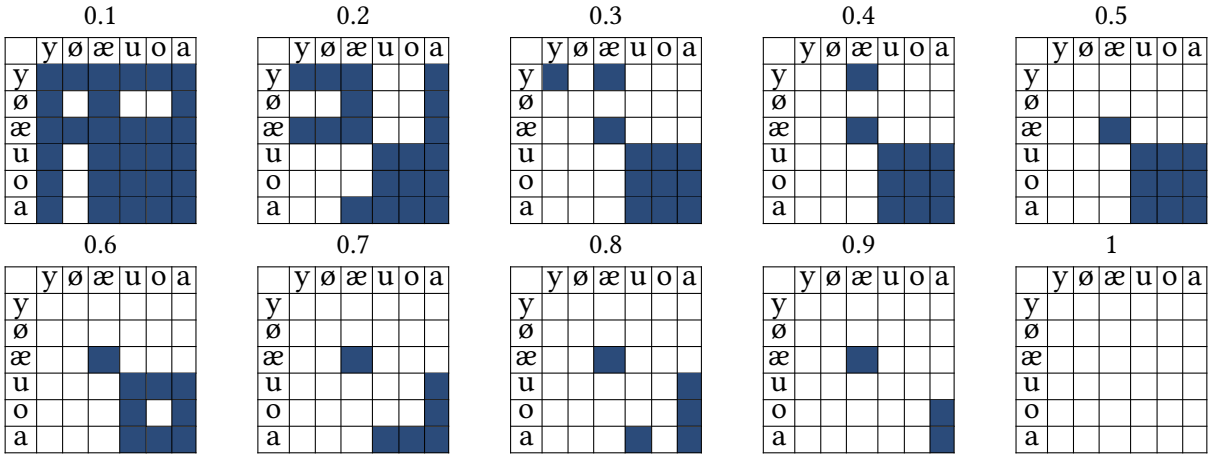


Figure 7.8: Learned Finnish phonotactic constraints (neutral vowels are omitted). Each individual subplot corresponds to a distinct $\theta_{\max}$ value, from 0.1 to 1, in phonotactic learning. Shaded cells indicates vowel two-factors accepted by the learned phonotactic grammar.

Figure 7.9 shows the outcomes of alternation learning. Maximal and minimal generalizations achieved comparable result, both outperforms the segment-based generalizations, except when $\theta_{\max} = 0.2$. When disharmonic two-factors are eliminated by the phonotactic grammar at $\theta_{\max} = 0.3$, the learned alternation grammar can achieve 96.08% accuracy in predicting the surface forms in the testing phase.

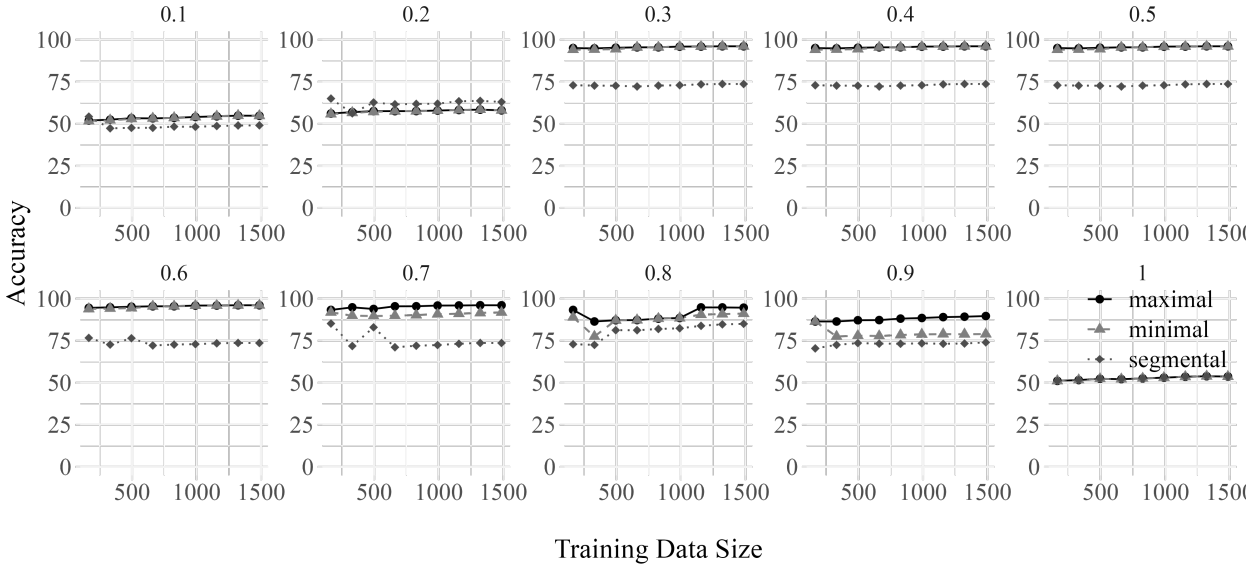


Figure 7.9: Finnish test result; each subplot corresponds to a distinct $\theta_{\max}$ value, from 0.1 to 1, in phonotactic learning; x-axis corresponds to the size of the training data; y-axis corresponds to the accuracy in predicting test data.

Table 7.8 shows the learned alternation grammar by the proposed learner at $\theta_{\max} = 0.3$. Most alternations in Finnish vowel harmony are learned, except the O $\rightarrow$ [-back]/[-back]___, corresponding to O $\rightarrow$ ø /[æ, ø, y]___ in segment-based generalizations. This is because two-factors and alternations related to ø is extremely rare in the training data. As previously discussed, these examples align with experimental evidence indicating that harmony patterns involving front rounded vowels are less represented and are consequently more likely to be inconsistently acquired or applied.

Segment-Based	Minimal Generalization	Maximal Generalization
A → æ / [æ, y]___	A → [+back]/[+back]___	A → [+back]/[+back]___
A → ɑ / [ɑ, u, o]___	A → [-back]/[-back]___	A → [-back]/[-back]___
U → u / [ɑ, u, o]___	U → [+back]/[+back]___	U → [+back]/[+back]___
U → y / [y]___	U → [-back]/[-back, +round, +high, -low]___	U → [-back]/[-back]___
O → o / [ɑ, u]___	O → [+back]/[+back]___	O → [+back]/[+back]___

Table 7.8: Learned Finnish alternations by the Two-Stage Phonotactic-Alternation Learner ($\theta_{\max} = 0.3$)

Table 7.9 shows the errors in predicting surface forms of the testing data. 3/10 errors can be attributed to interaction with other phonological processes such as the avoidance of vowel clusters (*VVV), and the raising conditioned by high front vowel or palatal glide ɑ → o /_ [i, j]. 7/10 errors can be traced back to compounds, loanwords, and lexicalized sequences. The learning model predicts that, learners will overextend the learned phonological alternations to these lexical forms before exposing to their observed forms. When vowel harmony patterns are not applied to these domains, front vowels [e, æ] are usually inserted as the default, as mentioned in the beginning of this section.

UR	Predicted	Observed	Note
haudu-tA-ttU	haudu-ta-ttu	haudu-te-ttu	Lexicalized *haudute* "stew"
haudu-tA-tU-lA	haudu-ta-tu-la	haudu-te-tu-la	Lexicalized *haudute* "stew"
budjeti-riih-i-stA	budjetti-riih-i-sta	budjetti-riih-i-stæ	French loanword "budget"
huolto-tyø-i-Vn	huolto-tyø-i-in	huolto-tø-i-hin	*VVV
perintø-raha-i-llA	perintø-raha-i-lla	perintø-raho-i-lla	ɑ → o /_ [i, j]
peru-nA-kaupa-i-stA-perunɑ	peru-na-kaupa-i-sta-perunɑ	peru-na-kaupo-i-sta-perunɑ	ɑ → o /_[i, j]
ilta-lehdi-ssA	ilta-lehdi-ssɑ	ilta-lehde-ssæ	Compound *ilta-lehdi* "evening-paper"
loppu-kiiree-ssA	loppu-kiiree-ssɑ	loppu-kiiree-ssæ	Compound "end-rush > in a hurry"
piirros-filmi-ssA	piirros-filmi-ssɑ	piirros-filmi-ssæ	Compound "drawing-film"
vaihto-piste-i-tA	vaihto-piste-i-tɑ	vaihto-piste-i-tæ	Compound "exchange-point"

Table 7.9: Errors in predicted Finnish surface forms by the Two-Stage learner (maximal; $\theta_{\max} = 0.3$; 90% MorphoChallenge); all surface forms are transcribed in IPA.

To summarize, the two-stage learner successfully acquires the phonotactic and alternation grammars from the training data, and achieved remarkable accuracy in predicting observed SRs in the testing data. The alternation learning performs at its best when the learned phonotactic grammar successfully filters out disharmonic two-factors, while not penalizing too many two-factors.

7.4 Summary

The current study achieves three objectives: (1) obtaining a learning model that induces human-like generalizations in phonotactic and alternation learning; (2) demonstrate the link between phonotactic and alternation learning; (3) predicting unseen test data.

When phonotactic grammar is too restricted, it causes undergenerlization in alternation learning; when it is unrestricted, it causes overgeneralization in alternation learning.

The current study serves as a baseline for more sophisticated learning models that integrate phonotactic and alternation learning....

CHAPTER 8

DISCUSSION OF THE TWO-STAGE PHONOTACTIC-ALTERNATION LEARNER

This section discusses the topics emerge from the proposal of Two-stage Phonotactic-Alternation learner. The current study achieves three objectives: (1) obtaining a learning model that induces human-like generalizations in phonotactic and alternation learning; (2) demonstrate the link between phonotactic and alternation learning; (3) predicting unseen test data. The case studies showed that, when phonotactic grammar is too restricted, it causes *undergeneralization* in alternation learning; when it is unrestricted, it causes *overgeneralization* in alternation learning as it fails to eliminate lexical exceptions from the input data.

8.1 Limitations and Future Directions in Test Data

The current study does not directly evaluate the phonotactic learning results against nonce word acceptability test data. Although the current study acknowledge this limitation, Dai (2023) has shown the capability of Exception-Filtering phonotactic learner in approximating gradient acceptability judgments in behavioral experiments, and a future direction is to collect acceptability judgments from children to validate the phonotactic learning results in the current study.

The current study does not evaluate the alternation learning results in *wug* test data from adult speakers (Berko, 1958), primarily due to the variability of adult behavioral data. For example, Hayes and Londe (2006) and Hayes et al. (2009) demonstrated high variability in Hungarian vowel harmony patterns in the wug tests of adult Hungarian speakers aged 18-75 years. Future research could include stochastic rules as per Albright and Hayes (2003b), which involves assigning probabilities to multiple productive rules based on their statistical support, thereby incorporating a wider range of linguistic variations.

It is worth noting that, the phonotactic and morphophonological knowledge acquired by children and adults is likely distinct, with children typically being exposed to a smaller size (<1k words) of input data across languages (Fenson et al., 1994; Hart and Risley, 1995; Bornstein et al., 2004; Szagun et al., 2006; Belth, 2023a). Children also display a higher rate of OVERREGULARIZATION in the application of morphophonological rules (Marcus et al., 1992; Yang, 2016).

8.2 Predictions of Language Development and Language Change

Previous acquisition studies have shown that phonotactic learning precedes alternation learning in child language acquisition: phonotactic learning initiates at an earlier stage and can evolve independently of morphological information. Before the age of 8 months, infants are already sensitive to nonlocal phonotactic patterns based on vowel harmony (Hohenberger et al., 2016), even when their native languages do not exhibit harmony patterns (Mintz et al., 2018; Sundara et al., 2022). In contrast, alternation learning is delayed and relies on the prior learning of a lexicon and morphological paradigms in order to track the mapping between different UR-SR forms (Hayes, 2004; Chong, 2017). Infants' sensitivity to alternation in perception is observed around 12 months in various languages (White et al., 2008; Skoruppa et al., 2013; Sundara et al., 2021), while relatively accurate production of morphophonological alternations does not occur until approximately 2.5 years of age (Smith, 1973; Marcus et al., 1992).

Moreover, the proposed learning algorithm in this dissertation has diachronic predictions (Kiparsky, 1965). The loss of productivity of phonological patterns is often associated with the presence of lexical exceptions, which are particular stems or other morphemes that idiosyncratically fail to undergo an alternation (Hayes, 2009). The relation between exceptionality and productivity has been a central topic in language acquisition (Yang, 2016).

The Two-stage learner predicts that, when the learned phonotactic grammar fails to eliminate exceptional sequences, the phonological processes cannot be learned, and they will become unproductive in loanwords and nonce words. For example, vowel harmony patterns across lan-

guages are on a continuum of productivity: Korean < Hungarian < Finnish < Turkish (Mahonen 2011:2). Jo (2024) and Jun et al. (2024) both argued that historically robust Korean vowel harmony lose productivity as they lose phonotactic support observed in the evolving lexicon.

8.3 Stochastic Constraint-based Frameworks

The current proposal models alternations as discrete, rule-based UR-SR mappings, which serve as a mathematical characterization that underlies any computational models of phonological acquisition, including stochastic, constraint-based frameworks.

Formally, a stochastic grammar is a set of weighted constraints/rules that predicts a probabilistic distribution over a set of possible SRs, given a UR, which is particularly convenient in explaining variations in linguistic data (Zuraw, 2000; Boersma and Hayes, 2001; Albright and Hayes, 2003b; Goldwater and Johnson, 2003; Hayes and Wilson, 2008). The concept of a stochastic phonological knowledge is supported by the phenomenon of FREQUENCY MATCHING observed in various experimental settings (Ernestus and Baayen, 2003; Hayes and Londe, 2006; Hayes et al., 2009), also known as GRADIENT PRODUCTIVITY (Moore-Cantwell and Pater, 2016; Shi and Emond, 2023). Participants were asked to produce novel words (i.e., the Wug test) or to make a perceptual response (e.g., rating tasks) to novel word stimuli, and their responses gradient match the frequencies of these patterns in their native language.

The current understanding of stochastic grammar and frequency matching, predominantly observed in adults, is challenged by recent studies in language acquisition showing the difference between adults and children's regularization. Hudson Kam and Newport found that adults' performance at the test phase matched the distribution of irregular patterns in their training input, but children regularized, generalizing beyond (or filtering out) the inconsistent input (Hudson Kam and Newport, 2005, 2009). Moreover, Austin et al. (2022) found that younger children regularize more than older ones. Furthermore, Shi and Emond (2023) observed that 14-month-olds do not exhibit gradient productivity in the generalizations of artificial word-order rules.

On the one hand, these findings suggest that future research to validate stochastic grammar should focus on children's language acquisition, considering these developmental differences. On the other hand, these findings can be interpreted as a developmental trajectory of the ability of frequency matching as the age (data size) grows. A future direction is to model such progression.

Constraint-based frameworks such as Optimality Theory (Prince and Smolensky, 1993) suggest that alternations are governed by the same principles as phonotactics, where the ranking of markedness over faithfulness constraints results in both static restrictions on surface forms and alternations adhering to those restrictions. Extending the Two-stage learner to constraint-based grammars requires non-trivial modification to ensure a finite candidate set of possible SRs for a given UR.

8.4 Refining Phonotactic Learning With Morphophonological Evidence

Many phonotactic learning models presuppose that learners have *a priori* knowledge of a threshold value (e.g., $\theta_{\max}$) that distinguishes between phonotactically well-formed and ill-formed sequences. In practice, this threshold value is commonly *a posterior* set by the analyst based on the predicative accuracy on testing data (Hayes and Wilson, 2008; Adriaans and Kager, 2010; Dai, 2023). However, the variability of this threshold across languages challenges the notion that it is an innate knowledge of learners. This approach raises a critical question regarding the phonotactic learning in children: if such a threshold is not innate, how do learners discern when their phonotactic knowledge is sufficiently accurate, especially in the absence of explicit negative evidence when learning from unlabelled, morphologically-agnostic data?

Insights from this paper's case studies suggest a potential answer: the input-output mappings in alternation learning can offer indirect negative evidence to refine the phonotactic learning threshold. Phonotactic learning is deemed *approximately correct* if it facilitates effective alternation learning. For example, without examining the blick test data, the Two-stage learner can determine $\theta_{\max} = 0.4$ achieves a higher accuracy in predicting new data in Turkish MorphoChal-

lenge (Figure 7.4b). Variations in learned phonotactic generalizations are acceptable as long as they enable the phonological learner to accurately discern input-output mappings.

Moreover, Hohenberger et al. (2016) showed that 6-month-old Turkish infants preferred listening to harmonic over disharmonic stimuli whereas 10-month-olds preferred listening to disharmonic over harmonic stimuli in longitudinal studies using a preferential listening paradigm. They argued that infants extract the general, regular, harmonic pattern and filter out irregular, disharmonic data. This regular-to-irregular shift is also shown in other acquisition studies, where children gradually develop sensitivity to irregular forms (Hudson Kam and Newport, 2005; Austin et al., 2022), as discussed above.

Combining the experimental evidence with the computational modeling on real-world corpora in the case studies, the current study proposes the following hypothesis to refine the procedure of phonotactic learning in the Two-stage learner: (1) Initialization phase: infants start learning phonotactics and alternations with a small amount of input data, and only pay attention to statistically overrepresented harmonic patterns.The phonotactic learning will start with a low $\theta_{\max}$, such as 0 that only penalize unattested sequences, and allow all attested sequences.

(2) Refining phase: when the input data size grows, children starts to acquire morphological information, and become more aware of the statistically underrepresented, disharmonic sequences in the data, and filter out these disharmonic tokens in alternation learning. This can be realized in Two-stage learner by adjusting $\theta_{\max}$ to minimize the contrast between the observed and predicted surface forms of stem + affixes.

(3) Convergence phase: when children grow even older, they encounter less and less novel stems and suffixes, both phonotactic and alternation grammar gradually converges.

This procedure is consistent the case studies examined in the current paper, and future studies are required to further test this hypothesis in other datasets.

8.5 Testing the Hypothesis Space

The current study employs a hypothesis space based on TSL_k languages, which are well-suited for deducing a broad array of local and nonlocal phonotactic patterns. The current study provides the foundation for testing the the hypothesis space involves using realistic corpora, and can be extended to superior yet *restrictive* structures.

Hayes and Wilson (2008) falsified their baseline hypothesis space, the SL_k languages, since it failed to learn nonlocal phonotactics in Shona input data. As a result, they switched to TSL_k languages. Similarly, the current paper utilizes the TSL_k hypothesis space as its baseline, which could also be empirically falsified if a better hypothesis space found to outperform the learning algorithm, when all other things being equal.

The TSL_k languages were not falsified in the current study because the learner was successful in learning local and nonlocal constraints. To date, the TSL_k hypothesis space has withstood tests against realistic corpora, although it has been questioned by isolated findings drawn from small, targeted datasets−−often gleaned from reference grammars or fieldwork data−−such as bidirectional ATR harmony patterns in Maasai and Turkana (McCollum et al., 2020). Future work can extend the proposed learning algorithms to alternative hypothesis spaces defined by other classes of formal languages.

CHAPTER 9

CONCLUSION

This research represents a significant step forward in several key areas: first, I've shown that phonological learning from noisy data can be achieved with a restrictive hypothesis space defined by formal language theory, esp. 2-TSL, and an exception-filtering mechanism based on type frequency. A deep insight from the exception-filtering learner is that learning is possible once the learner has access to high-quality, exception-free data. These are algorithm-independent insights that can benefit future learning models.

Moreover, this dissertation pioneers a "categorical grammar + exception-filtering mechanism" approach for learning categorical grammars from naturalistic input data with lexical exceptions. While the current study primarily focuses on the learning of grammars, it lays the groundwork for integrating learned grammars with extragrammatical factors to model behavioral data, and marks initial steps in reassessing the ability of categorical grammars in approximating human judgments.

Last but not the least, this dissertation models the link between phonotactic and alternation learning, and showed the prediction in language change driven by evolving lexicon. The learning algorithm predicts that the increased lexical exceptions cause the loss of phonotactic constraints and consequently the loss of productivity of related alternations.

BIBLIOGRAPHY

Adriaans, F. and Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, 62(3):311–331.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

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

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

Albright, A. and Hayes, B. (2002). Modeling english past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning*, pages 58–69.

Albright, A. and Hayes, B. (2003a). Learning nonlocal environments. Handout from the LSA meeting, Atlanta.

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

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

Altan, A. (2009). Acquisition of vowel harmony in turkish. *Dilbilim 35. Yıl Yazıları*, pages 9–26.

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

Alves, F. C. (2023). Categorical versus gradient grammar in phonotactics. *Language and Linguistics Compass*, 17(5):e12501.

Anderson, D. and Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63(2020):10.

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

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

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

Austin, A. C., Schuler, K. D., Furlong, S., and Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, 18(3):249–277.

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

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

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

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

Baković, E. (2007). A revised typology of opaque generalisations. *Phonology*, 24(2):217–259.

Belth, C. (2023a). Towards a learning-based account of underlying forms: A case study in turkish. *Proceedings of the Society for Computation in Linguistics*, 6(1):332–342.

Belth, C. A. (2023b). *Towards an Algorithmic Account of Phonological Rules and Representations*. PhD thesis, University of Michigan.

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

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

Berko, J. (1958). The child's learning of english morphology. *Word*, 14(2-3):150–177.

Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic inquiry*, 32(1):45–86.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. *Child development*, 75(4):1115–1139.

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

Burness, P. A., McMullin, K. J., Chandlee, J., Burness, P., and McMullin, K. (2021). Long-distance phonological processes as tier-based strictly local functions. *Glossa: a journal of general linguistics*, 6(1).

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Burzio, L. (2011). Derived environment effects. In *The Blackwell companion to phonology*, pages 1–26. Wiley Online Library.

Campbell, L. (1980). The psychological and sociological reality of finnish vowel harmony. In *Issues in vowel harmony*, pages 245–270. John Benjamins Amsterdam.

Chandlee, J. (2014). *Strictly local phonological processes*. University of Delaware.

Chandlee, J. and Jardine, A. (2021). Input and output locality and representation. *Glossa: a journal of general linguistics*, 6(1).

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

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

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

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

Chong, A. J. (2019). Exceptionality and derived environment effects: a comparison of korean and turkish. *Phonology*, 36(4):543–572.

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

Chong, J. A. (2017). *On the relation between phonotactic learning and alternation learning*. PhD thesis, UCLA.

Christensen, R. H. B. (2019). A tutorial on fitting cumulative link mixed models with clmm2 from the ordinal package. *Tutorial for the R Package ordinal https://cran. r-project. org/web/-packages/ordinal/Accessed*, 1.

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

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

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

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

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

Dai, H. (2023). A neo-trubetzkoyan approach to phonotactic learning in the presence of exceptions. Available at https://ling.auf.net/lingbuzz/007163.

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

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

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

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

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

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

Do, Y. and Yeung, P. H. (2021). Evidence against a link between learning phonotactics and learning phonological alternations. *Linguistics Vanguard*, 7(1):20200127.

Duncan, L. C. (2015). *Productivity of Finnish Vowel Harmony: Experimental Evidence*. University of Toronto (Canada).

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

Durvasula, K. and Liter, A. (2020). There is a simplicity bias when generalising from ambiguous data. *Phonology*, 37(2):177–213.

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

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

Emeneau, M. B. (1961). *Kolami: A Dravidian Language*. Annamalai University, Annamalainagar.

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

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, pages i–185.

Finley, S. (2010). Exceptions in vowel harmony are local. *Lingua*, 120(6):1549–1566.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368.

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

Frisch, S. A., Large, N. R., Zawaydeh, B., Pisoni, D. B., et al. (2001). Emergent phonotactic generalizations in english and arabic. *Typological studies in Language*, 45:159–180.

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

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

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

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

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

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

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

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

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

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

Gong, S. (2022). *The Obligatory Contour Principle Effects in Phonological Learning*. PhD thesis, University of Kansas.

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

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

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

Gorman, K. and Reiss, C. (2023). Maximal feature specification is feasible; minimal feature specification is not. lingbuzz/007296.

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

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

Hale, J. and Smolensky, P. (2006). Harmonic grammars and harmonic parsers for formal languages. In Smolensky, P. and Legendre, G., editors, *The Harmonic Mind, Volume 1: From Neural Computation to Optimality-Theoretic Grammar*, pages 393–416. MIT Press.

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

Halle, M. (1961). On the role of simplicity in linguistic descriptions. In *Proceedings of Symposia in Applied Mathematics: Structure of Language and its Mathematical Aspects*, volume 12, pages 89–94. American Mathematical Society.

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

Harrison, K. D. and Kaun, A. (2000). Pattern-responsive lexicon optimization. In *North East Linguistics Society*, volume 30, page 24.

Hart, B. and Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H Brookes Publishing.

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

Hayes, B. (2004). Phonological acquisition in optimality theory: The early stages. In Kager, R., Pater, J., and Zonneveld, W., editors, *Constraints in phonological acquisition*, pages 158–203. Cambridge University Press.

Hayes, B. (2008). *Introductory phonology*, volume 7. John Wiley & Sons.

Hayes, B. (2009). *Introductory phonology*, volume 7. John Wiley & Sons.

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

Hayes, B. (2016). Comparative phonotactics. In *Proceedings of the 50th meeting of the Chicago Linguistic Society*, pages 265–285.

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

Hayes, B., Siptár, P., Zuraw, K., and Londe, Z. (2009). Natural and unnatural constraints in hungarian vowel harmony. *Language*, pages 822–863.

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

Hayes, B. P. (2000). Gradient well-formedness in optimality theory[1]. *Optimality Theory: Phonology, syntax, and acquisition*, page 88.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley-Interscience.

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

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

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

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

Heinz, J. and Riggle, J. (2011). Learnability. *The Blackwell companion to phonology*, pages 1–25.

Hohenberger, A., Altan, A., Kaya, U., Tuncer, Ö. K., and Avcu, E. (2016). Sensitivity of turkish infants to vowel harmony: Preference shift from familiarity to novelty. In Ketrez, F. N. and Haznedar, B., editors, *The Acquisition of Turkish in Childhood*, pages 29–56. John Benjamins Publishing Company.

Hudson Kam, C. L. and Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, 1(2):151–195.

Hudson Kam, C. L. and Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66.

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

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

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

Iivonen, A. (1993). Paradigmaattisia ja syntagmaattisia näkökohtia lapsen foneettis-fonologisesta kehityksestä. In Iivonen, A., Lieko, A., and Korpilahti, P., editors, *Lapsen normaali ja poikkeava kielen kehitys*, volume 583 of *Suomalaisen Kirjallisuuden Seuran Toimituksia*, pages 34–77. Suomalaisen Kirjallisuuden Seura, Helsinki.

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

Itkonen, T. (1977). Huomioita lapsen äänteistön kehityksestä. *Virittäjä*, 81(3):279–308.

Iverson, G. K. and Wheeler, D. W. (1988). Blocking and the elsewhere condition. In *Theoretical morphology*, pages 325–338. Brill.

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

Jäger, G. and Rogers, J. (2012). Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598):1956–1970.

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

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

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

Jarosz, G. (2011). The roles of phonotactics and frequency in the learning of alternations. In *BU-CLD 35: Proceedings of the 35th annual Boston University Conference on Language Development*, pages 321–333.

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

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

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

Jo, J. (2024). Korean vowel harmony has weak phonotactic support and has limited productivity. *Phonology*.

Jun, J., Byun, H., Park, S., and Yee, Y. (2024). How tight is the link between alternations and phonotactics? *Phonology*.

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

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

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

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

Kabak, B. and Vogel, I. (2001). The phonological word and stress assignment in turkish. *Phonology*, 18(3):315–360.

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

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

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

Keane, J., Sehyr, Z. S., Emmorey, K., and Brentari, D. (2017). A theory-driven model of handshape similarity. *Phonology*, 34(2):221–241.

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

Kiparsky, P. (1965). *Phonological change.* PhD thesis, Massachusetts Institute of Technology.

Kiparsky, P. (1968). Linguistic universals and linguistic change. In Bach, E. and Harms, R. T., editors, *Universals in linguistic theory*, pages 170–202. Holt, Rinehart & Winston, New York.

Kiparsky, P. (1971). Historical linguistics. In Dingwall, W. O., editor, *A survey of linguistic science*, pages 576–642. University of Maryland Linguistics Program, College Park.

Kiparsky, P. (1993). Blocking in nonderived environments. In *Studies in lexical phonology*, pages 277–313. Elsevier.

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

Kuo, J. (2024). Phonological reanalysis is guided is guided by markedness: the case of malagasy weak stems. *Phonology*.

Kurimo, M., Virpioja, S., Turunen, V., and Lagus, K. (2010). Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95.

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

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

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

Leiwo, M. (1977). *Kielitieteellisiä näkökohtia viivästyneestä kielenkehityksestä*, volume 10 of *Studia Philologica Jyväskyläensia*. University of Jyväskylä, Jyväskylä.

Leiwo, M., Kulju, P., and Aoyama, K. (2006). The acquisition of finnish vowel harmony. *SKY Journal of Linguistics*, 19:149–161.

Lentz, T. O. and Kager, R. W. (2015). Categorical phonotactic knowledge filters second language input, but probabilistic phonotactic knowledge can still be acquired. *Language and speech*, 58(3):387–413.

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

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

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

Łubowicz, A. (2002). Derived environment effects in optimality theory. *Lingua*, 112(4):243–280.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

Mahonen, K. E. (2011). *The productivity of Finnish vowel harmony: An analysis of disharmonic words*. State University of New York at Buffalo.

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

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., and Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4):i–178.

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

Mayer, C. (2020). An algorithm for learning phonological classes from distributional similarity. *Phonology*, 37(1):91–131.

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

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

McCollum, A. G., Baković, E., Mai, A., and Meinhardt, E. (2020). Unbounded circumambient patterns in segmental phonology. *Phonology*, 37(2):215–255.

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

Mielke, J. (2008). *The emergence of distinctive features*. Oxford University Press.

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

Mintz, T. H., Walker, R. L., Welday, A., and Kidd, C. (2018). Infants' sensitivity to vowel harmony and its role in segmenting speech. *Cognition*, 171:95–107.

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

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

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

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

Nevins, A. (2011). Phonologically conditioned allomorph selection. In *The Blackwell companion to phonology*, volume 4, pages 2357–2382. Wiley-Blackwell New York.

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

Nişanyan, S. (2018). *Nişanyan Sözlük: Çağdaş Türkçenin Etimolojisi*. Liber Plus Yayınları, facsimile edition.

Nyman, A. and Tesar, B. (2019). Determining underlying presence in the learning of grammars that allow insertion and deletion. *Glossa: a journal of general linguistics*, 4(1):1–41.

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

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

O'Donnell, T. J., Goodman, N. D., and Tenenbaum, J. B. (2009). Fragment grammars: Exploring computation and reuse in language (tech. rep. no. mit-csail-tr-2009-013). *Massachusetts Institute of Technology*.

Pääkkönen, M. (1990). *Grafeemit ja konteksti – Tilastotietoja suomen yleiskielen kirjaimistosta "Graphemes and Context - Statistical Information on the Alphabet of Standard Finnish Language"*. SKS.

Padgett, J. (2002). Feature classes in phonology. *Language*, pages 81–110.

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

Pater, J. and Tessier, A.-M. (2006). L1 phonotactic knowledge and the l2 acquisition of alternations. In Slabakova, R., Montrul, S., and Prévost, P., editors, *Inquiries in Linguistic Development: Studies in Honor of Lydia White*, pages 115–131. John Benjamins Publishing Company, Amsterdam/Philadelphia.

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

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

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

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

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

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

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

Pizzo, P. (2015). *Investigating properties of phonotactic knowledge through web-based experimentation*. PhD thesis, University of Massachusetts Amherst.

Pizzo, P. and Pater, J. (2016). Does learning alternations affect phonotactic judgments? In *Proceedings of the Annual Meetings on Phonology*, volume 3.

Prickett, B. (2019). Learning biases in opaque interactions. *Phonology*, 36(4):627–653.

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

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

Pycha, A., Nowak, P., Shin, E., and Shosted, R. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In *Proceedings of the 22nd west coast conference on formal linguistics*, volume 22, pages 101–114. Cascadilla Press Somerville, MA.

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

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

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

Ringen, C. O. and Heinämäki, O. (1999). Variation in finnish vowel harmony: An ot account. *Natural Language & Linguistic Theory*, 17(2):303–337.

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

Rose, S. and King, L. (2007). Speech error elicitation and co-occurrence restrictions in two ethiopian semitic languages. *Language and Speech*, 50(4):451–504.

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

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

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

Shi, R. and Emond, E. (2023). The threshold of rule productivity in infants. *Frontiers in Psychology*, 14.

Shih, S. S. (2017). Constraint conjunction in weighted probabilistic grammar. *Phonology*, 34(2):243–268.

Skoruppa, K., Pons, F., Bosch, L., Christophe, A., Cabrol, D., and Peperkamp, S. (2013). The development of word stress processing in french and spanish infants. *Language Learning and Development*, 9(1):88–104.

Slobin, D. I. (1982). Universal and particular in the acquisition of language. *Language acquisition: The state of the art*, 57.

Smith, N. (1973). *The acquisition of phonology: A case study*. Cambridge University Press.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic inquiry*, 27(4):720–731.

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

Sundara, M., White, J., Kim, Y. J., and Chong, A. J. (2021). Stem similarity modulates infants' acquisition of phonological alternations. *Cognition*, 209:104573.

Sundara, M., Zhou, Z., Breiss, C., Katsuda, H., and Steffman, J. (2022). Infants' developing sensitivity to native language phonotactics: A meta-analysis. *Cognition*, 221:104993.

Szagun, G., Steinbrink, C., Franik, M., and Stumper, B. (2006). Development of vocabulary and grammar in young german-speaking children assessed with a german language development inventory. *First language*, 26(3):259–280.

Szeredi, D. (2016). *Exceptionality in vowel harmony*. PhD thesis, New York University.

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

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

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

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

Vago, R. I. (1988). Vowel harmony in finnish word games. In Van der Hulst, H. and Smith, N., editors, *Features, Segmental Structure and Harmony Processes, Part II*, pages 185–205. Foris Publications, Dordrecht.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vitevitch, M. S. and Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, 9(4):325–329.

Wang, Y. and Hayes, B. (2022). Learning underlying representations: Expectation-maximization and the kk-hierarchy. Poster presented at the Annual Meeting on Phonology, UCLA, Los Angeles, CA, October 2022.

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

Whang, J. and Adriaans, F. (2017). Phonotactics and alternations in the acquisition of japanese high vowel reduction. In *Proceedings of the Boston University Conference on Language Development (BUCLD)*, pages 41–730.

White, K. S., Peperkamp, S., Kirk, C., and Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, 107(1):238–265.

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

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

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

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

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

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

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

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

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

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